Identification automatisée de marques descriptives sur des images de douilles de cartouche

par

Marie-Eve LE BOUTHILLIER

THÈSE PRÉSENTÉE À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION
DU DOCTORAT EN GÉNIE
Ph. D.

MONTRÉAL, LE 23 OCTOBRE 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE :

M. Luc Duong, directeur de thèse
Département de génie logiciel et TI à l'École de technologie supérieure

Mme Sylvie Ratté, codirectrice de thèse
Département de génie logiciel et TI à l'École de technologie supérieure

M. Mohamed Cheriet, président du jury
Département de génie des systèmes à l'École de technologie supérieure

M. Christopher Fuhrman, membre du jury
Département de génie logiciel et TI à l'École de technologie supérieure

Mme Rola Harmouche, examinatrice externe
Conseil National de Recherches Canada

ELLE A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 12 SEPTEMBRE 2025

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Cette thèse vient clore plusieurs années de travail, de recherche et de collaboration, au sein d'un environnement académique enrichissant. Elle a été rendue possible grâce au soutien et aux conseils de nombreuses personnes que je tiens à remercier très sincèrement.

Je remercie avant tout mon directeur de thèse, Luc Duong, et ma codirectrice, Sylvie Ratté, pour leur accompagnement, leur rigueur scientifique et leur disponibilité, tout au long des trois dernières années. Leurs commentaires toujours constructifs ont grandement contribué à la qualité de ce travail. Leurs expertises en apprentissage machine, en traitement de l'image, ainsi qu'en structure de manuscrit de thèse ont profondément enrichi cette recherche.

Je souhaite également exprimer ma reconnaissance à Alain Beauchamp, dont l'initiative et la vision ont été à l'origine de ce projet. Merci d'avoir cru en son potentiel — et en moi — dès les premières étapes de sa réalisation.

Un grand merci à mes collègues Lynne Hrynkiw et Alain Fidahoussen avec qui j'ai partagé, toutes les deux semaines, des discussions passionnées sur les projets d'intelligence artificielle (et parfois sur des sujets un peu moins académiques…). Ces échanges ont été une source constante d'inspiration et de motivation. Ils m'ont permis d'affiner mes idées et de faire avancer concrètement cette recherche.

Je tiens aussi à remercier les membres des laboratoires de recherche LIVE et LiNCS, pour nos rencontres hebdomadaires, où la discussion scientifique côtoyait souvent la bonne humeur. Ces échanges ont fréquemment été le carburant dont j'avais besoin pour relancer ma réflexion… et mon moral !

Un merci sincère aux experts qui ont accepté de participer anonymement à ce projet. Votre contribution précieuse, bien que discrète, a joué un rôle clé dans l'avancement de cette

recherche. Votre implication a renforcé la validité des résultats et témoigne de l'importance du dialogue entre théorie et pratique.

À Andrée, merci du fond du cœur pour ton soutien sans faille — même quand tu ne comprenais pas vraiment ce que je faisais. Tu as su, mieux que personne, trouver les mots (ou les silences) qu'il me fallait à chaque étape de ce parcours.

Je remercie LeadsOnline pour le financement du projet.

En dernier lieu, je tiens à remercier l'École de technologie supérieure d'avoir fourni un cadre académique stimulant dans lequel cette recherche a pu s'épanouir.

Identification automatisée de marques descriptives sur des images de douilles de cartouche

Marie-Eve LE BOUTHILLIER

RÉSUMÉ

Les scènes de crimes impliquant des armes à feu doivent être analysées attentivement en vue de présenter un ensemble de preuves lors d'un procès. Il s'avère utile pour les enquêtes de pouvoir associer une arme à feu avec une scène de crime. Lors du tir d'une cartouche, les mécanismes de l'arme à feu laissent des marques descriptives sur la balle et la douille. Lorsque des douilles sont retrouvées sur une scène de crime, ces marques sont analysées dans un laboratoire. Ces marques peuvent être des caractéristiques de classe, qui sont communes aux armes à feu d'une même famille; ou des caractéristiques individuelles, qui sont théoriquement uniques et permettent de relier un spécimen à une arme à feu précise. Avant de comparer les caractéristiques individuelles pour proposer des correspondances, un triage manuel est généralement effectué en considérant les caractéristiques de classe, afin d'éviter de calculer inutilement les comparaisons avec des spécimens produits par des familles d'armes différentes. Une analyse automatique permettrait d'assister les techniciens lors de la saisie de l'information, en suggérant les catégories présentes.

L'objectif principal de cette thèse est de déterminer une méthode d'apprentissage machine permettant d'identifier des marques microscopiques présentes sur des images de douilles de cartouche. La première contribution consiste à classifier les marques selon sept catégories: parallèle, arche, hachure, circulaire, granulaire, lisse et inconnu. Nous avons entraîné et évalué des modèles multiclasses, multiétiquettes, et binaires. L'évaluation du modèle multiétiquette ENB3 sans augmentation présente une mesure F1 de 54,47%, et une perte de 0,36. Les évaluations des modèles binaires sans augmentation présentent des valeurs entre 70,00% et 93,00% pour la mesure F1, et des valeurs entre 0,22 et 2,01 pour la fonction de perte. Lors de ces expérimentations, nous avons remarqué des incohérences dans les étiquettes de la vérité de terrain et nous suggérons un ré-étiquetage des données.

La seconde contribution consiste à évaluer des regroupements d'images de douilles. Ces regroupements ont été créés par des algorithmes traditionnels et un réseau profond de clustering, à partir des caractéristiques descriptives des images extraites par les réseaux entraînés à l'objectif précédent. Nous avons observé que les algorithmes favorisant la création de clusters de formes et de tailles variées semblent mieux adaptés à nos données, et nous avons apprécié la technique de clustering flou, qui calcule un degré d'appartenance des points à chaque groupe. Nous suggérons des travaux futurs combinant ces deux approches.

La dernière contribution étudie l'accord interobservateur entre six observateurs humains et les méthodes d'apprentissage profond supervisé. Nous avons observé des coefficients Kappa moyens variant de faibles pour les catégories granulaire et lisse, à élevés pour la catégorie circulaire. Nous avons ensuite construit un ensemble de vérité de terrain de 1000 échantillons vérifié par deux experts. Nous avons repris les expérimentations multiétiquettes avec cet

ensemble, sans augmentation. Nous avons observé une amélioration, avec une courbe d'entraînement atteignant 90,00% d'exactitude. Lors de l'évaluation, nous avons observé une amélioration de la moyenne pondérée pour la mesure F1 à 77,12%. Nous croyons que ce modèle démontre de meilleures capacités pour la classification des images de douilles et nous suggérons d'améliorer l'ensemble de vérité de terrain en ajoutant des échantillons supplémentaires, particulièrement dans les catégories plus rares.

En dernière analyse, il ressort de cette thèse que l'apprentissage profond pourrait être utilisé afin d'identifier automatiquement des marques descriptives sur des images de douilles de cartouche. Éventuellement, des modèles de classification pourraient être inclus dans des systèmes d'identification automatisée. En plus d'améliorer les performances des algorithmes de comparaison en réduisant les temps de calcul, ils pourraient fournir de l'information aux enquêteurs promptement, ce qui pourrait permettre de faire avancer les enquêtes en cours plus rapidement. Les modèles devront toutefois être améliorés afin d'inclure les catégories plus rares, ainsi que certaines exceptions (telles que des marques produites à d'autres moments que lors de l'utilisation de l'arme). Nous recommandons de poursuivre la recherche avec les classificateurs binaires tout en augmentant l'ensemble de vérité de terrain vérifié.

Mots-clés : Apprentissage profond, multiétiquette, classification, regroupement, clustering, défauts de surface

Automated identification of descriptive marks on images of cartridge cases

Marie-Eve LE BOUTHILLIER

ABSTRACT

Crime scenes involving firearms need to be carefully analyzed to present a body of evidence at trial. It is useful for investigations to be able to associate a firearm with a crime scene. When a cartridge is fired, the mechanisms of the firearm leave descriptive marks on the bullet and the cartridge case, and when cartridge cases are found at a crime scene, these marks are analyzed in a laboratory. These marks may be class characteristics, which are common to firearms in the same family; or individual characteristics, which are theoretically unique and enable a specimen to be linked to a specific firearm. Before individual characteristics are compared to propose matches, manual sorting is generally carried out by considering class characteristics, to avoid unnecessary calculation of comparisons with specimens produced by different gun families. An automatic analysis would assist technicians when entering information, by suggesting the categories present.

The main objective of this thesis is to determine a machine learning method for identifying microscopic marks present in images of cartridge cases. The first contribution consists in classifying the marks according to seven categories: parallel, arch, crosshatch, circular, granular, smooth and unknown. We trained and evaluated multiclass, multilabel and binary models. Evaluation of the ENB3 multilabel model without augmentation shows an F1 measure of 54.47%, and a loss of 0.36. The evaluations of the binary models without augmentation show values between 70.00% and 93.00% for the F1 measure, and values between 0.22 and 2.01 for the loss function. During these experiments, we noticed inconsistencies in the ground-truth labels and suggest relabelling the data.

The second contribution consists in evaluating clusters of cartridge case images. These clusters were created by traditional algorithms and a deep clustering network, based on the descriptive features of the images extracted by the networks trained in the previous objective. We observed that algorithms favouring the creation of clusters of various shapes and sizes seemed better suited to our data, and we appreciated the fuzzy clustering technique, which calculates a degree of membership of points to each group. We suggest future work combining these two approaches.

The final contribution study inter-observer agreement between six human observers and supervised deep learning methods. We observed average Kappa coefficients ranging from low for the granular and smooth categories, to high for the circular category. We subsequently built a ground truth set of 1,000 samples verified by two experts. We repeated the multilabel experiment with this set, without data augmentation. We observed an improvement, with a training curve reaching 90.00% accuracy. In the evaluation, we observed an improvement in the weighted average for the F1 measure to 77.12%. We believe that this model demonstrates better capabilities for the classification of cartridge case images,

and we suggest improving the ground truth set by adding additional samples, particularly those in the more uncommon categories.

In summary, this thesis shows that deep learning could be used to automatically identify descriptive marks on images of cartridge cases. Eventually, classification models could be included in automated identification systems. As well as improving the performance of matching algorithms by reducing computation times, they could provide information to investigators promptly, which could help to advance ongoing investigations faster. The models will, however, need to be improved to include rarer categories, as well as certain exceptions (such as marks produced at times other than when the weapon was being used). We recommend continuing research with binary classifiers while increasing the verified ground truth set.

Keywords: Deep learning, multi-labelling, classification, clustering, surface defects

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AFTE	Association of firearm and tool mark examiners *Association des examinateurs d'armes à feu et de marques d'outils*
ANOVA	Analysis of variance *Analyse de la variance*
AUC	Area under the ROC curve *Surface sous la courbe ROC*
BF	Breech face *Face de la culasse*
CAM	Class activation mapping *Cartographie de l'activation des classes*
CCD	Charge-coupled device (camera) *Dispositif à couplage de charge (caméra)*
CFmax	Maximum cross-correlation function *Fonction de corrélation croisée maximale*
CH	Calinsky Harabasz index *Indice de Calinsky Harabasz*
CMC	Congruent matching cells *Cellules de correspondance congruente*
CNN	Convolutional neural network *Réseau neuronal convolutif*
CPT	Conformal prediction theory *Théorie de la prédiction conforme*
DBSCAN	Density-based spatial clustering of applications with noise *Regroupement spatial des applications avec bruit, basé sur la densité*
DCN	Deep clustering network *Réseau de clustering profond*
DCNN	Deep convolutional neural networks *Réseau neuronal convolutif profond*
DEC	Deep embedded clustering *Clustering intégré profond*
DES	Dynamic ensemble selection *Sélection dynamique d'ensembles*
DWT	Discrete wavelet transform *Transformée discrète d'ondelettes*

ELM	Extreme learning machine *Machines d'apprentissage extrême*
ELW	Entropy-guided loss weighting *Pondération de la perte guidée par l'entropie*
FC	Fully connected *Entièrement connecté*
FCM	Fuzzy C-means *C-moyens flou*
FCN	Fully convolutional neural network *Réseau neuronal entièrement convolutif*
FP	Firing pin *Percuteur*
FROVOCO	Fuzzy rough OVO combination *Combinaison OVO floue et rugueuse*
GA	Genetic algorithm *Algorithme génétique*
GAP	Global average pooling *Mise en commun des moyennes générales*
GLCM	Gray-level co-occurrence matrix *Matrice de cooccurrence des niveaux de gris*
GradCam	Gradient-weighted class activation mapping *Cartographie d'activation de classe pondérée par le gradient*
HDBSCAN	Hierarchical density-based spatial clustering of applications with noise *Regroupement spatial hiérarchique des applications avec bruit, basé sur la densité*
HOG	Histogram of oriented gradient *Histogramme du gradient orienté*
IBIS	Integrated ballistic identification system *Système d'identification balistique intégré*
ICP	Iterative closest point *Point le plus proche itératif*
InfoGAN	Information maximizing generative adversarial network *Réseau adversatif génératif maximisant l'information*
KNN	K-nearest neighbours *K-voisins les plus proches*

LBP	Local binary patterns *Motifs binaires locaux*
LDA	Linear discriminant analysis *Analyse discriminante linéaire*
LPQ	Local phase quantization *Quantification de phase locale*
LR	Likelihood ratio *Rapport de vraisemblance*
LoG	Laplacian of Gaussian *Laplacien Gaussien*
MAcc	Mean accuracy *Moyenne de la précision*
MAUC	Mean of area under the curve *Moyenne de l'aire sous la courbe*
MFM	Mean of F-measure *Moyenne de la mesure F1*
MLP	Multi-layer perceptron *Perceptron multicouche*
NIBIN	National integrated ballistics information network *Réseau national intégré d'information balistique*
NIST	National institute of standards and technology *Institut national des normes et de la technologie*
NLP	Natural language processing *Traitement du langage naturel*
OPTICS	Ordering points to identify the clustering structure *Ordonner les points pour identifier la structure de regroupement*
PCA	Principal component analysis *Analyse en composantes principales*
PSNR	Peak Signal to Noise Ratio *Rapport signal sur bruit maximal*
ROC	Receiver operating characteristic *Caractéristique fonctionnelle du récepteur*
ROI	Region of interest *Région d'intérêt*

RANSAC	Random sample consensus *Consensus sur l'échantillon aléatoire*
RMSE	Root Mean Square Error *Erreur quadratique moyenne*
SIFT	Scale invariant feature transform *Transformation des caractéristiques invariantes à l'échelle*
SSIM	Structural Similarity *Indice de similarité structurelle*
SURF	Speeded up robust features *Accélération des caractéristiques robustes*
SSL	Semi-supervised learning *Apprentissage semi-supervisé*
SVD	Singular value decomposition *Décomposition en valeurs singulières*
SVM	Support vector machine *Machine à vecteur de support*
SMO	Sequential minimal optimization algorithm *Algorithme d'optimisation minimale séquentielle*
TSNE	T-distributed stochastic neighbour embedding *Intégration de voisins stochastiques distribués*
UMAP	Uniform manifold approximation and projection *Approximation et projection d'un manifold uniforme*
UQI	Universal Image Quality Index *Indice universel de qualité d'image*
VGG	Visual geometry group *Groupe de géométrie visuelle*
ViT	Vision transformer *Transformeur de vision*
WCL	Weighted cross-entropy loss *Perte d'entropie croisée pondérée*
WK	Wavelet kernel *Noyau d'ondelettes*

INTRODUCTION

Selon Amnistie internationale, la violence liée aux armes à feu provoque chaque jour la mort de plus de 600 personnes, représentant environ 71 % des homicides à l'échelle mondiale. On estime que 85 % du milliard d'armes en circulation appartiennent à des particuliers. Cette situation touche particulièrement les quartiers urbains à faibles revenus, souvent marqués par une criminalité élevée et des services policiers insuffisants ou non conformes aux standards internationaux en matière de droits de la personne. Les survivants subissent des séquelles durables, tant physiques que psychologiques, et les violences frappent de manière disproportionnée les populations déjà vulnérables, notamment les hommes et les garçons issus de communautés défavorisées, ainsi que les personnes racisées ou marginalisées (Amnesty International, [s d]). Au-delà des victimes directes, ces actes affectent également les témoins et les proches des victimes, impactant le bien-être général de la société. Un traitement rapide et rigoureux des enquêtes, visant à reconstituer les faits et à fournir des preuves devant les tribunaux pour sanctionner les responsables, peut apporter un soulagement significatif aux victimes et à leurs familles, tout en contribuant à prévenir l'apparition de nouvelles victimes grâce à l'arrestation des coupables avant toute récidive.

0.1 Contexte

Dans le cadre d'une enquête criminelle, l'analyse des éléments matériels présents sur une scène de crime constitue une étape fondamentale. Les balles et les douilles sont particulièrement importantes, car elles peuvent être directement reliées à l'arme utilisée. Cette identification repose sur l'examen des marques microscopiques laissées par l'arme, qui fonctionnent comme une empreinte balistique unique. On distingue alors deux types de caractéristiques : celles de classe ou de sous-classe, susceptibles d'être partagées par plusieurs armes à feu, et celles dites individuelles, considérées comme théoriquement uniques.

Lorsqu'une douille est récupérée, elle est transmise à un laboratoire spécialisé, où le processus débute par une inspection visuelle et par l'enregistrement de ses caractéristiques dans un système automatisé. Avant l'analyse balistique, toutes les traces pertinentes, telles que l'ADN ou les empreintes digitales, sont prélevées. La douille est ensuite soumise à une analyse au moyen de dispositifs d'imagerie de pointe, permettant la production d'images 2D de haute résolution ainsi que de données topographiques 3D. Ces techniques offrent des conditions d'observation précises et reproductibles. L'automatisation de ces étapes contribue à limiter l'influence des biais humains et favorise la standardisation des résultats. Dès cette étape, l'identification de certaines marques de classe pourrait fournir des indications pertinentes aux enquêteurs.

Les images obtenues sont ensuite transmises à un réseau central doté d'une base de données, qui les compare à celles des douilles de la même famille de calibres et calcule un score de similitude pour chaque paire. À ce stade, un tri préliminaire basé sur les marques de classe pourrait permettre de limiter les comparaisons aux échantillons présentant des marques similaires, éliminant ainsi les cas non pertinents. En réduisant le bruit dans les données, cette préclassification pourrait raccourcir les temps de calcul et améliorer les performances des algorithmes de comparaison.

Les résultats finaux de ces systèmes ont une portée juridique significative, puisqu'ils peuvent constituer des éléments de preuve présentés devant les tribunaux. Une correspondance confirmée établit qu'une cartouche a été tirée par une arme spécifique, contribuant ainsi à l'arrestation ou à la condamnation d'un suspect. À l'inverse, une absence de correspondance permet d'exclure une arme particulière, démontrant qu'elle n'a pas produit les marques et pouvant ainsi soutenir la défense d'une personne.

Traditionnellement, les experts en balistique et en identification d'outils examinaient les échantillons visuellement à l'aide d'un microscope de comparaison, à la recherche de marques permettant de confirmer ou d'éliminer l'utilisation d'une arme spécifique. Aujourd'hui, l'informatique s'est intégrée à plusieurs disciplines de la médecine légale, telles

que la comparaison d'empreintes digitales et l'analyse de preuves ADN. La science de la balistique n'a pas fait exception : depuis plusieurs années, des méthodes de traitement d'image et d'apprentissage automatique assistent les enquêteurs dans l'analyse des éléments retrouvés sur les scènes de crime, contribuant ainsi à faire progresser les enquêtes. (Gerules, Bhatia et Jackson, 2013).

À ce jour, peu de recherches se sont penchées sur l'utilisation de méthodes basées sur l'apprentissage profond pour l'identification des marques de classe sur des images de douilles de cartouches. Dans ce projet, des techniques d'apprentissage automatique et d'apprentissage profond seront mises en œuvre afin d'identifier automatiquement ces marques de classe. À notre connaissance, il n'existe actuellement aucune procédure automatisée dédiée à ce type de classification. Ce projet pourra être intégré à un processus de comparaison automatisé dans le but d'améliorer les performances d'appariement des douilles. De plus, la possibilité de découvrir de nouveaux critères de classification pourrait également contribuer à l'optimisation des algorithmes existants.

0.2 Problématique

L'identification balistique constitue un enjeu central des enquêtes criminelles, puisqu'elle permet d'établir un lien entre une arme à feu et une pièce à conviction, comme une balle ou une douille. Or, ce processus repose encore largement sur l'expertise humaine et l'examen visuel des marques microscopiques laissées par l'arme. Cette dépendance à l'évaluation subjective entraîne une variabilité interobservateur et soulève des questions quant à la reproductibilité et la standardisation des résultats.

Dans ce contexte, l'émergence des approches fondées sur l'apprentissage profond ouvre de nouvelles perspectives pour automatiser certaines étapes de l'analyse balistique et renforcer son objectivité. La problématique est donc la suivante : dans quelle mesure ces méthodes peuvent-elles transformer l'analyse des douilles en un processus plus fiable, plus rapide,

permettant d'examiner une quantité toujours plus importante de données, et reposant moins sur l'interprétation subjective de l'expert ?

0.3 Questions de recherche

Pour y répondre, trois questions de recherche seront explorées :

1. Est-ce qu'il est possible d'améliorer le processus d'identification balistique grâce à une assistance automatisée de la détection des marques de classe ?
2. L'apprentissage non supervisé est-il en mesure de révéler des régularités significatives dans les marques balistiques, indépendamment d'un étiquetage préalable des données ?
3. Quelle est la variabilité interobservateur, et comment se compare-t-elle avec les méthodes automatiques?

CHAPITRE 1

REVUE DE LITTÉRATURE

1.1 Introduction

Cette revue de littérature se divise en trois parties. La première partie traite du domaine médico-légal et de la balistique traditionnelle. Elle présente quelques notions de la théorie de l'identification des marques d'outils et discute des marques typiquement présentes sur les douilles de cartouche. La seconde partie discute de la criminalistique informatique. Elle présente quelques études publiées sur la balistique, en s'attardant sur la comparaison, la détection de la région d'intérêt et la classification. La troisième partie jette un coup d'œil sur des études dans des domaines différents de celui de la balistique, qui se sont intéressées à la classification ou qui se sont tournées vers l'apprentissage non supervisé afin de détecter des marques sur des pièces variées, principalement avec des techniques de clustering.

1.2 Domaine médico-légal

Cette section présente un bref résumé du domaine médico-légal et de la science de la balistique.

1.2.1 Balistique

Les échantillons balistiques sont des balles et des douilles provenant d'une cartouche tirée par une arme à feu. Ces échantillons contiennent des marques caractéristiques pouvant fournir des indices importants lors des enquêtes criminelles et pourraient éventuellement constituer des éléments de preuve qui seront présentés devant les tribunaux (Gerules, Bhatia et Jackson, 2013). Les marques caractéristiques présentes sur les échantillons deviennent une empreinte balistique, qui permet aux spécialistes de l'identification balistique d'obtenir des informations sur l'arme qui a été utilisée pour tirer la cartouche. Certaines marques permettent même de déterminer si une cartouche a été tirée par une arme à feu spécifique,

résultant dans l'appréhension ou la condamnation d'un criminel; ou encore, d'exclure une arme à feu spécifique avec la conclusion qu'elle n'a pas pu produire les marques, prouvant ainsi l'innocence d'une personne (Gerules, Bhatia et Jackson, 2013 ; Kara, 2016 ; Pisantanaroj *et al.*, 2020 ; Riva et Champod, 2014 ; Zheng *et al.*, 2014).

Afin de se prononcer, le spécialiste de l'identification balistique se base sur la détection, la reconnaissance et la correspondance entre les caractéristiques présentes sur les échantillons (Riva *et al.*, 2017). Puisque cette tâche est complexe, une expertise est nécessaire afin de distinguer les similitudes dans les empreintes. Avant l'arrivée du microscope de comparaison dans les années 20, la mémoire des examinateurs ainsi que l'alignement manuel de photographies sur papier étaient utilisés afin d'évaluer les correspondances entre les échantillons. Le microscope de comparaison (Figure 1.1) a contribué grandement à l'accélération du processus d'identification en permettant aux experts d'aligner et d'observer simultanément les surfaces agrandies de deux échantillons (Kara, 2016 ; Zheng *et al.*, 2014).

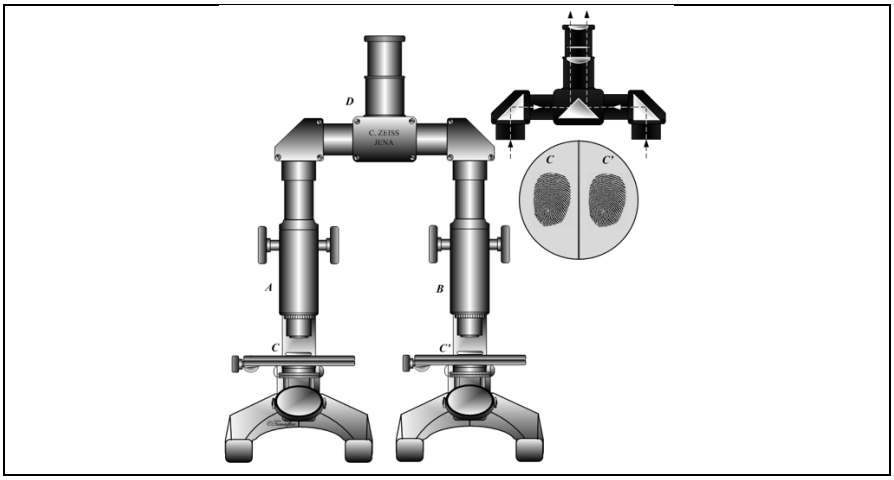


Figure 1.1 Microscope de comparaison
Tirée de (Tamasflex, 2010)

Le fondement scientifique de cette identification repose sur deux principes. Le premier principe indique que les marques produites par une arme sont cohérentes et reproductibles d'un tir à l'autre. Le second principe indique que les marques individuelles sont uniques et différenciables entre deux armes à feu (Fischer et Vielhauer, 2015). Autrement dit, un

ensemble particulier de marques présentes sur un échantillon balistique compose une signature balistique dont l'unicité constitue la base de la théorie de l'identification (Riva et Champod, 2014 ; Song *et al.*, 2014). Il est à noter qu'il est reconnu que les marques retrouvées sur les échantillons balistiques provenant d'une même arme à feu ne sont pas toujours complètement identiques, malgré le principe de reproductibilité (des effets d'usure ou de corrosion pourraient altérer l'apparence de ces marques avec le temps, par exemple). De plus, bien que plusieurs travaux de recherche soutiennent ces deux fondements d'unicité et de reproductibilité, plusieurs estiment qu'ils ne sont pas encore entièrement démontrés et que des travaux supplémentaires sont nécessaires afin de corroborer leur validité (Riva et Champod, 2014).

En outre, certaines études mentionnent des contestations concernant les techniques de comparaison manuelles des échantillons balistiques et la validité de cette science est mise en question. On lui reproche entre autres l'absence de tests rigoureux suivant une méthode scientifique, sa nature subjective inhérente reposant sur la formation et l'expérience précise d'un examinateur particulier, ainsi que la possibilité de différentes formes de biais pouvant affecter le jugement humain de l'expert. Le manque de critères et de normes objectives définissant des procédures d'identification ainsi que des méthodes d'évaluations est aussi souvent énoncé (Gerules, Bhatia et Jackson, 2013 ; Kara, 2016 ; Riva *et al.*, 2020 ; Riva et Champod, 2014). En 2009, un rapport émis par les Académies Nationales a recommandé l'élaboration de critères objectifs pour l'identification des marques d'outils (Zheng *et al.*, 2014). Selon certains, le domaine émergent de la criminalistique informatique pourrait contribuer à réduire la subjectivité en offrant une analyse objective ou en émettant une probabilité statistique (Gerules, Bhatia et Jackson, 2013 ; Riva *et al.*, 2020 ; Riva et Champod, 2014).

1.2.2 Marques d'outils

L'identification des marques d'outils est une discipline de la médecine légale qui s'intéresse à déterminer si un outil particulier a été utilisé pour produire une marque sur un autre objet

(Morris *et al.*, 2017). Les marques d'outils sont définies comme étant une déformation produite par le contact de deux surfaces, dont la plus dure est appelée l'outil. Suivant cette définition, une arme à feu est considérée comme un outil qui transfère des marques sur les balles et les douilles tirées (Gerules, Bhatia et Jackson, 2013 ; Zheng *et al.*, 2014 ; Kudonu *et al.*, 2022).

On distingue principalement deux sortes de marques d'outils. Les marques microscopiques striées (striated) sont habituellement produites lorsqu'un outil, dont le mouvement est approximativement parallèle à une surface, est glissé le long de cette surface. Les marques microscopiques imprimées (impressed) ou les empreintes sont produites lorsqu'un outil, dont le mouvement est approximativement perpendiculaire à une surface, frappe ou appuie sur cette surface. Dans les échantillons balistiques, les marques striées présentent une topographie de surface ressemblant à des lignes parallèles (striae) et se retrouvent principalement sur les balles. Les empreintes présentent une topographie négative de l'outil et se retrouvent principalement sur les douilles (Morris *et al.*, 2017 ; Zheng *et al.*, 2014).

Les examinateurs tentent d'identifier deux catégories de marques, retrouvées autant sur les marques striées que sur les empreintes (Morris *et al.*, 2017). Les caractéristiques de classe ou de sous-classe, pouvant être partagées par plusieurs armes à feu; et les caractéristiques individuelles, théoriquement uniques (AFTE, 2013 ; Gambino *et al.*, 2011). Dans un article publié en 2017, Riva *et al.* mentionnent l'existence de directives générales visant à aider les examinateurs lors de l'identification de ces caractéristiques, ainsi que plusieurs travaux de recherche portant sur l'origine de ces marques lors de différents processus de fabrication (Riva *et al.*, 2017). Zheng *et al.* mentionnent des études de cas portant sur des outils fabriqués séquentiellement ayant démontré que les topographies de surface étaient uniques et identifiables pour tous les outils testés (Zheng *et al.*, 2014). Lors de l'examen d'un échantillon balistique, les experts commencent généralement par l'identification et la comparaison des caractéristiques de classe et s'attardent ensuite aux caractéristiques individuelles.

1.2.2.1 Caractéristiques de classe et de sous-classe

Les caractéristiques de classe sont communes à tous les outils d'un certain type ou d'un même modèle. Leurs formes permettent ainsi de déterminer le type ou le modèle de l'outil utilisé pour produire la marque, mais ces marques ne permettent pas d'identifier un outil spécifique. Si cette caractéristique s'applique à un petit nombre d'outils seulement, par exemple une série d'outils fabriqués de façon séquentielle, on parle alors de caractéristiques de sous-classe (Morris *et al.*, 2017 ; Zheng *et al.*, 2014).

1.2.2.2 Caractéristiques individuelles

Les caractéristiques individuelles sont des marques apparaissant sur la surface d'un outil. Il peut aussi s'agir de variations, d'irrégularités, d'imperfections ou de dommages impartis lors de la production, de l'utilisation ou de l'usure de l'outil. Ces marques sont spécifiques à un outil particulier et elles permettent d'identifier l'outil distinct qui a produit la marque examinée (Kudonu *et al.*, 2022 ; Zheng *et al.*, 2014 ; Gambino *et al.*, 2011).

1.2.2.3 Théorie de l'identification en rapport avec les marques d'outils

La base de la science de la balistique repose sur la théorie de l'identification en rapport avec les marques d'outils établie par l'AFTE, qui stipule que l'on peut conclure *« à une origine commune lorsque les contours uniques de la surface de deux marques d'outils sont en concordance suffisante »* [traduction libre] (AFTE, 2013). La concordance s'obtient en comparant les caractéristiques individuelles correspondantes entre deux échantillons. C'est-à-dire en comparant les hauteurs, les profondeurs, les largeurs, les courbures et les relations spatiales des différents motifs de contours de surface, par exemple des pics, des crêtes ou des sillons. Une *concordance suffisante* signifie que la concordance entre les deux marques d'outils évaluées dépasse les concordances démontrées avec des marques produites par des outils différents et qu'elle est cohérente avec la concordance démontrée avec des marques produites par le même outil. Essentiellement, que la qualité et la quantité de la concordance entre les caractéristiques individuelles réduisent la possibilité qu'un autre outil ait pu

produire la marque à une quasi-impossibilité (AFTE, 2013 ; Gambino *et al.*, 2011 ; Riva *et al.*, 2017).

Gambino et al. suggèrent que ces marques pourraient être considérées comme des modèles mathématiques, permettant ainsi l'utilisation de différentes méthodes numériques (pour des tâches de classification, par exemple) afin d'attacher des mesures quantitatives objectives plutôt qu'un simple terme de *concordance suffisante* (Gambino *et al.*, 2011).

1.2.3 Douille de cartouche

Une cartouche de fusil est constituée d'un projectile (la balle) et d'un contenant (la douille). La douille, habituellement composée d'un alliage de laiton, contient la balle ainsi que le matériel nécessaire au tir, tel que le propulseur. Dans les armes à feu modernes, les poudres sans fumée à base de nitrocellulose ou de composés organiques de même nature ont remplacé la poudre noire qui était utilisée comme propulseur à l'origine. À l'extrémité de la douille se trouve un rebord qui permet d'abord de positionner la cartouche dans la chambre de l'arme et ensuite d'extraire et d'éjecter la douille une fois que la balle est tirée. Finalement, une amorce située sur la base de la douille contient une charge explosive qui enflamme le propulseur lorsqu'elle est frappée par le percuteur de l'arme à feu, voir Figure 1.2 (Gerules, Bhatia et Jackson, 2013).

Le processus de tir d'une arme à feu laisse des marques singulières striées sur la balle, causées par des rayures situées à l'intérieur du canon du pistolet; alors que le percuteur, la face de la culasse du canon, l'extracteur et l'éjecteur *(firing pin, breech face, extractor and ejector)* laissent des marques caractéristiques imprimées sur la douille. Des marques striées peuvent aussi être produites sur la douille lors de l'insertion de la cartouche *(chambering)* ou de l'alimentation du chargeur *(magazine feed)*. Le nombre et le type de marque varient selon le modèle de l'arme à feu (Gerules, Bhatia et Jackson, 2013 ; Song *et al.*, 2014).

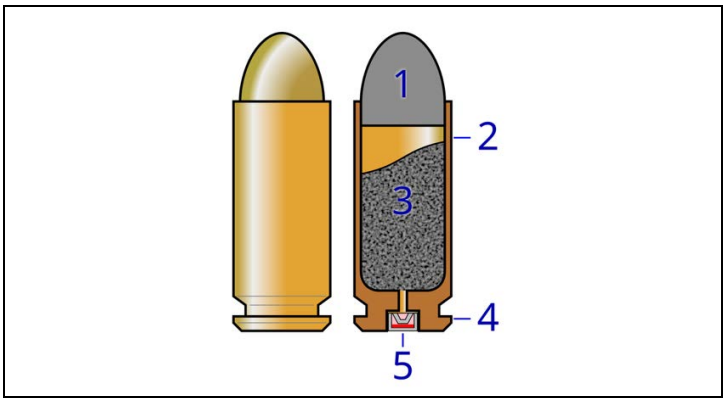


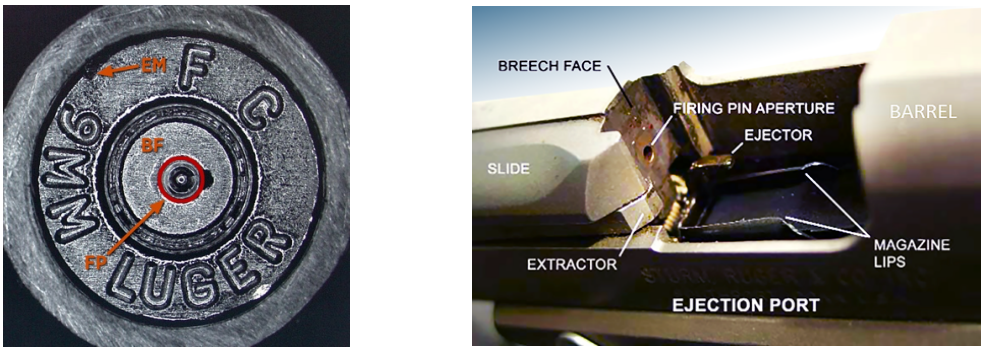
Figure 1.2 Principales sections d'une cartouche :
1 – balle, 2 – douille, 3 – propulseur,
4 – bordure (*rim*), 5 – amorce (*primer*)
Tirée de (Glrx, 2021)

1.2.3.1 Caractéristiques de classe sur une douille

Le cachet du fabricant *(headstamp)* est souvent étampé sur la tête de la douille, ainsi que le type et le calibre de la munition. Dans certains cas, la forme de la marque du percuteur peut aussi être classée selon différentes catégories. Par exemple : circulaire, elliptique, rectangulaire, double ou annulaire. Ces marques peuvent permettre d'identifier le type, le calibre ou le fabricant de l'arme, mais sans pouvoir isoler une arme spécifique (Gerules, Bhatia et Jackson, 2013).

1.2.3.2 Caractéristiques individuelles sur une douille

Sur les douilles de cartouche, trois zones peuvent exhiber des marques d'outils provenant de différents mécanismes de l'arme à feu. La Figure 1.3 présente ces trois principales zones de marques ainsi que les composantes de l'arme à feu responsable de les avoir produites.



a *b*

Figure 1.3 La marque de la face de la culasse BF, l'empreinte du percuteur FP et la marque de l'éjecteur EM (a), composantes de l'arme à feu (b)
Tirées de (NIST, 2012 ; NIST, 2015)

- **Percuteur**

  Pour entamer la mise à feu de la balle, le percuteur frappe la capsule de la cartouche afin de faire exploser l'amorce. La topographie négative de la surface du percuteur est alors imprimée sur la douille, créant ainsi l'empreinte du percuteur. La forme, la position, les dimensions et la profondeur de cette marque sont distinctives selon le modèle de l'arme à feu. On retrouve plusieurs formes de percuteur et cette empreinte est souvent distinctive lors de l'observation des marques (Kara, 2016 ; Zheng *et al.*, 2014).

- **Face de la culasse**

  La face de la culasse est une surface plate sur laquelle la cartouche est appuyée. Lors du tir de la balle, la force explosive qui permet d'enflammer le propulseur pousse la douille dans la direction opposée, contre la face de la culasse. La topographie négative de cette surface s'imprime alors sur la tête de la douille, créant ainsi la marque de la face de la culasse. Pour l'identification, ces empreintes sont considérées comme étant des marques d'outils fiables (Zheng *et al.*, 2014 ; Zhu *et al.*, 2022).

- **<u>Éjecteur</u>**

  Les armes semi-automatiques éjectent les douilles de cartouche une fois la balle tirée. À cette fin, un extracteur tire d'abord la douille hors de la chambre et un éjecteur frappe ensuite le bord arrière de la douille. Cette dernière action imprime la topographie négative de la surface de l'éjecteur, créant ainsi l'empreinte de l'éjecteur. Puisqu'elles sont ainsi éjectées automatiquement, plusieurs douilles sont souvent retrouvées sur une scène de crime (Pisantanaroj *et al.*, 2020 ; Zheng *et al.*, 2014).

Les procédés de fabrication et de traitements des pièces de la face de la culasse et du percuteur, ainsi que les matériaux distincts utilisés, sont suffisamment différents pour présumer de l'unicité des traces laissées par ces pièces sur les douilles (Morris *et al.*, 2017). Des études ont tenté d'établir dans quelle mesure les marques produites par la culasse sont affectées par les propriétés physiques de la douille, par exemple la dureté des matériaux ou la composition et l'épaisseur de la capsule d'amorce *(primer cap)*. Ces études donnent un bon point de départ dans l'identification en réduisant les combinaisons possibles, mais elles demeurent contestées dans la littérature : aucun consensus clair ne définit comment ces propriétés affectent les marques (Gerules, Bhatia et Jackson, 2013).

Il est important de préciser que l'empreinte du percuteur et la marque de la face de la culasse sont produites à deux moments distincts lors du tir de la balle. L'empreinte du percuteur est partiellement créée avant l'inflammation du propulseur, et sa morphologie peut ensuite être modifiée par d'autres mécanismes de l'arme, par exemple des marques de traînée du percuteur peuvent être produites lors de l'éjection. L'empreinte de la face de la culasse est créée lors de l'augmentation de la pression dans la chambre. Étant donné que la position de la douille peut varier entre les instants de la frappe FP et de la collision BF, l'alignement relatif des deux marques peut varier d'un échantillon à l'autre. Par conséquent, ces empreintes sont généralement séparées avant d'effectuer une analyse balistique informatique (Riva et Champod, 2014).

1.3 Identification automatique pour le jumelage des échantillons balistiques

Depuis les années 80, plusieurs chercheurs se sont tournés vers l'informatique afin de tenter d'améliorer la science de l'identification balistique en proposant différentes analyses quantitatives objectives sur la correspondance entre deux échantillons. Le développement des équipements d'acquisition d'images de surface par topographie 3D, l'avancement de la technologie de traitement des images, ainsi que les travaux de recherche sur les algorithmes automatisés de comparaison permettant d'établir une correspondance ont permis d'automatiser et d'accélérer les processus d'acquisition et de jumelage des échantillons balistiques, améliorant ainsi l'identification des armes à feu. Plusieurs systèmes automatisés, généralement composés d'un microscope optique, d'une procédure d'extraction de signature, d'une station d'analyse de signature et d'un logiciel de corrélation, ont été développés. Par exemple, le système IBIS développé au début des années 1990 en Amérique (UEFT, [s d]) et la base de données de preuves balistiques NIBIN *« National Integrated Ballistics Information Network »*, mis en place en 1999 (NIBIN, [s d]), fournissent une estimation quantitative brevetée des correspondances les plus probables (Kudonu *et al.*, 2022). Il est tout de même important de spécifier que toute correspondance ou tout jumelage, obtenu par un outil informatique ou confirmé par un expert humain, constitue une concordance statistiquement viable, plutôt qu'une correspondance absolue (Gerules, Bhatia et Jackson, 2013 ; Goodfellow, Bengio et Courville, 2016 ; Song *et al.*, 2014 ; Tai et Eddy, 2018 ; Zheng *et al.*, 2014).

1.3.1 Images topographiques 3D pour les échantillons balistiques

Depuis les années 30, des images 2D en tons de gris sont utilisées dans les enquêtes pour examiner les échantillons balistiques. Grâce aux méthodes d'imagerie complexe des technologies plus récentes, il est possible aujourd'hui de produire des images 3D de la topographie de surface des échantillons balistiques. Contrairement aux photographies 2D qui offrent une mesure indirecte de la topographie de la surface en évaluant les variations de pente et les ombres et qui sont ainsi impactées par la position de la lumière et la réflexion, la topographie 3D permet une représentation mesurée d'une surface. La microscopie virtuelle

(l'étude numérique automatisée de la topographie de surface 3D) permet de recueillir des caractéristiques de surface étendues, basées sur les informations de topographie et pouvant être combinées avec des méthodes de reconnaissance de formes. Dans la dernière décennie, plusieurs travaux ont été effectués afin de concevoir des algorithmes de comparaisons se basant sur des images 3D. Les résultats obtenus ont montré que les mesures topographiques 3D permettent d'obtenir une bonne précision ainsi qu'un niveau élevé de reproductibilités (Fischer et Vielhauer, 2015 ; Gerules, Bhatia et Jackson, 2013 ; Kara, 2016 ; Kudonu *et al.*, 2022 ; Riva et Champod, 2014 ; Tai et Eddy, 2018 ; Zheng *et al.*, 2014).

1.3.2 Algorithmes de comparaison entre deux images de douille

Plusieurs études portant sur les marques d'échantillons balistiques ont proposé des solutions d'automatisation des algorithmes de comparaison afin de présenter une valeur de correspondance quantitative et objective (Zheng *et al.*, 2014). Kara avait utilisé une méthode de classement par corrélation *(correlation ranking method)* afin d'assigner des valeurs numériques exprimant la similarité et la différence entre deux images de marques sur une douille. Leur système peut être utilisé pour différencier les types et les modèles de pistolets (Kara, 2016).

Gambino *et al.* mentionnent l'absence de système universellement accepté pouvant générer une valeur de similitude numérique objective et permettant de corroborer indépendamment les correspondances entre les douilles évaluées par des observations visuelles. Les auteurs proposent un système permettant d'évaluer objectivement des correspondances entre des marques d'outils striées et l'outil responsable de les avoir administrées. L'étude teste 58 images topographiques 3D de douilles contenant des marques de cisaillement de l'amorce provenant du tir de quatre armes à feu du même modèle. Après avoir filtré les images pour faire ressortir les caractéristiques fondamentales sous forme de profils d'ondulation *(waviness profiles)*, des méthodes d'apprentissage machine ont été utilisées afin d'associer les profils aux armes à feu correspondantes. Une analyse en composantes principales (PCA) est d'abord utilisée pour réduire la dimensionnalité des caractéristiques. Ensuite, une

machine à vecteur de support (SVM) est entraînée par apprentissage supervisé afin de classifier les profils d'ondulation. Finalement, les niveaux de confiance sont estimés en utilisant une prédiction conforme (CPT). Bien que cette étude ne porte que sur un nombre restreint d'échantillons et un seul modèle d'arme à feu, les auteurs recommandent l'utilisation de leur méthode dans un contexte légal telle une salle d'audience (Gambino *et al.*, 2011).

Un certain nombre d'études sur l'appariement des marques de la face de la culasse s'est intéressé à la méthode des cellules de correspondance congruente (CMC) proposée initialement en 2013 pour des topographies 3D. Depuis, la méthode a été revisitée par le National Institute of Standards and Technology (NIST) en vue d'adapter son application aux images 2D afin qu'elle soit compatible avec les bases de données locales et nationales, constituées majoritairement d'images 2D (Tai et Eddy, 2018). Au fil des années, différents travaux ont contribué à l'amélioration de la technique. L'idée de base consiste à identifier les zones de corrélation valides et à éliminer les zones de corrélation invalides en divisant l'image en plusieurs cellules; puis à comparer les cellules valides en utilisant une fonction de corrélation croisée (Song *et al.*, 2014 ; Zhu *et al.*, 2022).

En 2015, une étude a proposé et évalué deux caractéristiques invariantes à la rotation, destinées à être utilisées dans un système de détection et de reconnaissance de formes à partir de données topographiques de douilles afin de déterminer automatiquement des correspondances entre des empreintes de percuteurs. Puisque la base des cartouches est circulaire, l'utilisation de caractéristiques invariantes à la rotation pourrait permettre d'analyser les images de douilles sans nécessiter un alignement de la rotation; réduisant ainsi l'interaction avec l'utilisateur ainsi que la possibilité d'erreurs. Les caractéristiques invariantes à la rotation proposées combinent des valeurs de quantités de lignes multiples, circulaires ou droites avec des évaluations statistiques et ont été convenablement baptisées : MCP *(Multiple-Circle-Path)* et MAP *(Multiple-Angle-Path)*. Une machine à vecteur de support (SVM) se charge de la classification, utilisant les vecteurs de caractéristiques individuelles en entrée et une technique de validation croisée stratifiée dix plis *(folds)*.

L'ensemble de données utilisé contient 72 douilles provenant de six armes à feu de trois modèles différents, avec trois types de munitions. Trois types d'expérimentation ont été effectués, mais la section des résultats contient peu d'explications. Les auteurs affirment que leur méthode dépasse deux autres études de l'état de l'art, mais en se référant seulement aux résultats présentés dans les articles originaux et sans avoir testé ces méthodes sur leur ensemble de données. Il est donc difficile de se situer sur les gains réels de leur technique (Fischer et Vielhauer, 2015).

En 2018, une étude propose une méthode publiée en logiciel libre entièrement automatisée, pour comparer les marques de la face de la culasse. La méthode développée utilise des images 2D en niveaux de gris, mais les auteurs affirment qu'elle est tout aussi utilisable sur des topographies 3D. Les marques sont sélectionnées automatiquement et les images sont traitées afin de réduire les effets de symétrie circulaire. La concordance est évaluée empiriquement en utilisant la fonction de corrélation croisée maximale (CCFmax) et la probabilité de correspondance aléatoire sur des comparaisons par paires. Les tests ont été effectués à partir d'un nombre restreint de douilles provenant de tirs d'essai : 108 douilles, provenant de 12 pistolets différents, appartenant à trois modèles sélectionnés et utilisant trois marques de munitions. Les résultats démontrent une amélioration générale de la précision et les auteurs mentionnent que la réduction des effets de symétrie circulaire est efficace sur les scores de similitudes des échantillons non correspondants alors que peu d'impact est observé sur les scores de similitudes des échantillons correspondants (Tai et Eddy, 2018).

Dans une étude publiée en 2020, les auteurs ont évalué les performances d'un système de comparaison automatique visant à réduire la subjectivité des évaluations balistiques en attribuant une valeur de poids aux correspondances et aux dissemblances des marques d'outils identifiées sur des douilles. Ce système permet de comparer ainsi les empreintes du percuteur, les marques de la face de la culasse, ainsi qu'une combinaison de ces deux marques. Pour déterminer cette valeur de poids, le rapport de vraisemblance (LR) est calculé à partir des mesures topographiques 3D et d'un modèle statistique bidimensionnel. Trois mesures de similarités quantitatives sont utilisées : l'indice de corrélation, une mesure basée

sur la distance Euclidienne et une mesure basée sur les vecteurs normaux à la surface. Afin d'aligner automatiquement les images, la méthode Simplex (un algorithme basé sur l'optimisation) a remplacé l'algorithme ICP utilisé dans des études précédentes. Les résultats démontrent que le système possède certaines limites, particulièrement avec les petits ensembles de données. Par contre, les auteurs suggèrent que ce système pourrait aider les examinateurs dans un contexte légal, en renforçant leurs décisions (Riva *et al.*, 2020). Dans une étude préalable utilisant un système similaire, Riva avait démontré que les rapports de vraisemblances (LR) sont indicatifs de la réalité dans un contexte légal (Riva et Champod, 2014). Dans une autre étude préalable, Riva avait utilisé un système similaire pour évaluer l'impact des caractéristiques de classe et de sous-classe sur la mesure LR. Principalement, l'étude cherchait à déterminer si la présence de caractéristiques de classe et de sous-classe sur une douille, communes à plusieurs spécimens balistiques ayant été tirés par un même modèle d'arme à feu, pouvait influencer la force d'une valeur de correspondance déterminée automatiquement. Pour y arriver, des douilles provenant de 13 armes à feu distinctes, mais du même modèle ont été comparées; cependant, l'explication de la procédure de test demeure floue. Les résultats supportent que l'impact des marques de classe et de sous-classe soit faible, mais qu'il puisse devenir considérable lorsque la proportion d'armes à feu partageant les caractéristiques de la classe ou de la sous-classe dépasse 40% de l'ensemble testé (Riva *et al.*, 2017).

En 2022, une étude a proposé une méthode basée sur la similarité d'images pour l'examen balistique d'images de douilles, provenant de tirs d'essai de 20 marques et modèles d'armes automatiques. Au total, 1000 images de cartouches 9 × 19 mm (Parabellum) et 7,65 × 17 mm (0,32 ACP) ont été comparées avec l'indice de similarité structurelle (SSIM), l'erreur quadratique moyenne (RMSE), le rapport signal sur bruit maximal (PSNR) et l'indice universel de qualité d'image (UQI) pour 4 types d'évaluations: la marque de la face de la culasse, la marque du percuteur, la marque de l'éjecteur, ainsi qu'une évaluation combinée. Les résultats montrent que SSIM serait la méthode la plus efficace pour analyser la correspondance des échantillons. Les auteurs envisagent également d'explorer l'apprentissage profond dans le futur, afin d'améliorer le taux de réussite et de développer

une plateforme d'analyse balistique complète. Bien que le terme « classification » soit utilisé dans cette étude, il ne semble pas correspondre à la notion classique de classification supervisée en apprentissage automatique. Il paraît plutôt désigner un classement ordonné des scores de similarité pour les paires de douilles. Les auteurs mentionnent aussi une étape de pré-sélection incluant la reconnaissance des marques caractéristiques, mais ne fournissent aucun détail à ce sujet.

Finalement, un nombre limité d'études a exploré l'analyse d'échantillons balistiques avec des méthodes utilisant l'apprentissage profond. Kudonu *et al.* ont évalué la capacité de cette approche à fournir des analyses objectives sur des balles en examinant des caractéristiques de classe, comme les marques de rayures. Les auteurs suggèrent des algorithmes d'apprentissage non supervisé pour l'analyse des spécimens balistiques, mais sans indiquer les raisons (Kudonu *et al.*, 2022). Toujours en 2022, une autre étude a utilisé Superpoint (un algorithme d'apprentissage profond autosupervisé) avec des images de la face de la culasse. Superpoint est un réseau neuronal entièrement convolutif (FCN) permettant d'extraire les caractéristiques des points d'intérêt dans une image, d'identifier les points clés et de proposer un descripteur local pour chacun d'entre eux. Le réseau est formé d'un encodeur de type VGG, d'un décodeur de points d'intérêt et d'un décodeur de descripteurs. Les appariements sont déterminés en utilisant les méthodes de correspondance par force brute, de correspondance par rapport à la distance et RANSAC. L'étude utilise 40 images de la face de la culasse, provenant de douilles tirées par dix armes à feu distinctes, mais manufacturées de façon séquentielle. La performance dépasse celle obtenue avec la méthode traditionnelle SIFT et les résultats montrent que le modèle Superpoint suivi d'une correspondance des caractéristiques peut être utilisé avec succès pour apparier des images de marques sur la face de la culasse (Zhu *et al.*, 2022).

1.3.3 Sélection de la région d'intérêt

Dans la plupart des études impliquant des échantillons balistiques, la détection et la sélection de la région d'intérêt ne sont pas toujours bien définies. Ces étapes sont parfois même

ignorées dans les rapports, on suppose alors qu'elles ont été faites à la main préalablement. Un article explique que dans de nombreuses études, la région d'intérêt est sélectionnée manuellement en positionnant un premier cercle entourant la région de l'amorce et un second cercle entourant l'empreinte du percuteur (Tai et Eddy, 2018).

Dans une étude publiée en 2011, une description de l'acquisition de l'image du cachet de la tête est suivie d'une explication sur la position des marques de cisaillement de l'amorce; pour ensuite sauter directement aux méthodes de prétraitement de l'image permettant de faire ressortir les caractéristiques de la marque d'outil (Gambino *et al.*, 2011). Dans le même ordre d'idées, une étude de 2014 mentionne découper les surfaces n'appartenant pas à la marque de la face de la culasse (la surface située à l'intérieur de la zone de l'empreinte du percuteur et celle située à l'extérieur de la zone de la marque de la face de la culasse), mais ne mentionne aucun détail sur les techniques de découpage automatique utilisées (Song *et al.*, 2014). Dans une autre étude publiée en 2014, les auteurs expliquent qu'ils doivent positionner manuellement la douille, afin que le plan de la coupelle d'amorce soit parallèle avec le plan horizontal, avant d'isoler la partie supérieure de la coupelle et de séparer les zones de la marque de la face de la culasse et de l'empreinte du percuteur avec un algorithme de segmentation automatique utilisant les vecteurs normaux à la surface, mais aucun détail supplémentaire n'est fourni sur cet algorithme (Riva et Champod, 2014). Une étude de 2015 mentionne détecter automatiquement le cercle extérieur du percuteur, sans donner plus de détails sur la technique. Les auteurs précisent que cette étape est effectuée pendant les prétraitements, et que les images sont ensuite annotées manuellement (Fischer et Vielhauer, 2015). Finalement, dans un article publié en 2020, les auteurs mentionnent que des étapes de prétraitement ont été utilisées pour segmenter les images, incluant la segmentation automatique de la coupelle d'amorce et la séparation automatique des différents types de marques telles que l'empreinte du percuteur et la marque de la face de la culasse, mais sans préciser les méthodes utilisées pour y parvenir (Riva *et al.*, 2020).

Dans la majorité des articles où elle est décrite, la détection automatique de la région d'intérêt est effectuée avec des méthodes traditionnelles de traitement de l'image,

principalement avec des méthodes de détection de contours telles que l'opérateur de Sobel et le détecteur de contours de Canny. Des opérations de segmentation sont aussi parfois utilisées, lorsqu'il est nécessaire d'isoler la région d'intérêt de l'image de fond. Des techniques de seuillage de l'image sont alors utilisées, produisant une image binaire affichant des pixels noir ou blanc et utilisée surtout pour effectuer des analyses de la forme. Gerules et al. mentionnent quelques études ayant utilisé les techniques de détection de contours de Sobel ou Canny, ainsi qu'une étude ayant utilisé les informations de hauteur de l'image accompagnées d'une transformée de Hough pour faire la segmentation de différentes sections de la douille (Gerules, Bhatia et Jackson, 2013).

Néanmoins, il existe quelques articles qui présentent une description détaillée de cette étape de sélection. Par exemple, dans une étude menée en 2018, les marques de la surface de la face de la culasse sont sélectionnées automatiquement grâce à des méthodes traditionnelles de traitement de l'image. Dans cet article, qui inclut plusieurs figures illustrant les différentes actions, les étapes de la sélection de la zone d'intérêt sont bien expliquées : la région de l'amorce est d'abord identifiée avec une combinaison d'opérations (filtre gaussien, égalisation de l'histogramme, remplissage par diffusion, dilatation et érosion). L'empreinte du percuteur est ensuite enlevée grâce à une combinaison d'opérations similaires, à laquelle s'ajoute une détection des contours de Canny. Cet ajout permet de conserver la forme réelle de la marque du percuteur, ce qui est intéressant puisque cette forme possède une valeur d'identification (Tai et Eddy, 2018).

1.3.4 Classification automatique des échantillons balistiques

Très peu d'études ont exploré la classification automatisée des marques sur des échantillons balistiques. Par exemple, Pisantanaroj *et al*. ont utilisé des méthodes d'apprentissage profond afin de classifier des images panoramiques de balles, selon huit catégories d'armes à feu. Afin de rendre le système portable, une méthode permettant de prendre les photographies panoramiques à partir d'un téléphone mobile a été développée. Trois architectures de réseau de neurones ont été préentraînées sur l'ensemble de données d'ImageNet en utilisant

l'apprentissage par transfert : DenseNet121, ResNet50 et Xception; avec une technique de validation croisée en cinq plis. La performance des différents modèles a été évaluée avec des courbes ROC et des mesures AUC et les valeurs de sensibilité globale, de spécificité et de précision ont été utilisées afin d'évaluer les résultats. L'ensemble de données testé contient 718 balles appartenant aux huit catégories d'armes à feu à l'étude, mais dans des proportions variées. Les auteurs spécifient que les échantillons ayant été tirés par la même arme à feu ont été volontairement placés dans le même pli pour l'entraînement des modèles. Les résultats démontrent que le modèle DenseNet offre la meilleure performance en général, ainsi que pour la catégorie d'arme la plus utilisée en Thaïlande, le pays d'origine de l'étude. Cependant, les résultats varient selon le type d'arme, et les explications tendent à devenir confuses en exprimant plusieurs mesures, itérations d'expérimentation et distinctions. Les auteurs expliquent que le déséquilibre des données ne semble pas avoir affecté les résultats, mais ils ajoutent qu'il s'agit tout de même d'une possibilité. Ils précisent avoir utilisé une analyse ANOVA équilibrée à deux facteurs afin d'identifier des différences entre les modèles et de détecter une interaction entre le modèle et le type d'arme à feu, mais les explications qui s'ensuivent ne sont pas claires. Ils concluent en exprimant le potentiel prometteur de leur système portable (Pisantanaroj *et al.*, 2020). Même si leur méthode demeure simple (une image complète soumise telle quelle à une architecture de réseau neuronal), les résultats semblent encourageants pour des classifications plus complexes telles que l'identification et la classification d'une ou plusieurs types de marques sur un échantillon.

1.4 Apprentissage machine pour la reconnaissance et la classification d'images

De nos jours, plusieurs domaines se tournent vers les techniques d'apprentissage machine pour résoudre des problèmes impliquant des tâches de vision par ordinateur. Certains chercheurs utilisent des méthodes traditionnelles de traitement de l'image pour extraire des caractéristiques utiles de l'image, et un algorithme d'apprentissage machine permet ensuite de les catégoriser. D'autres s'intéressent aux applications rendues possibles par l'évolution rapide de l'apprentissage profond, dont la capacité à égaler et même à surpasser les performances humaines pour les tâches de reconnaissance de formes a été démontrée à de

multiples reprises. En outre, l'accès à différentes options de logiciels libres ainsi que la disponibilité d'un vaste réseau de ressources facilitent grandement l'utilisation de ces techniques modernes (Byeon *et al.*, 2019).

1.4.1 Apprentissage non supervisé pour la détection d'anomalies

Face à l'identification de marques d'outils ou de défaut sur des surfaces, quelques études se sont tournées vers des méthodes d'apprentissage non supervisées, notamment des techniques de clustering. La détection d'anomalies est une tâche importante pour le contrôle de la qualité en industrie, et son automatisation est considérée comme un problème difficile. La petite taille des défauts par rapport à l'image évaluée, l'étendue de la variation de leur forme, l'arrière-plan et les différences d'éclairage sont autant de facteurs cités pour expliquer la difficulté de la tâche puisqu'ils complexifient l'extraction des caractéristiques et affectent la détection (Manimozhi et Janakiraman, 2019 ; Tan *et al.*, 2020 ; Tomczak, Mosorov et Sankowski, 2006).

Dans les études utilisant des méthodes de clustering sur des images pour identifier des défauts, les termes *segmentation*, *clustering* et *classification* sont parfois utilisés de façon interchangeable. La segmentation par texture consiste à identifier et à séparer les régions d'une image possédant des caractéristiques de texture distinctes et se divise principalement en deux étapes : l'extraction des caractéristiques de texture et leur classification. Plusieurs méthodes ont été utilisées avec succès pour l'extraction des caractéristiques de texture, par exemple : les statistiques d'histogramme, la matrice de cooccurrence, l'autocorrélation et le modèle binaire local. Les approches basées sur des modèles, telles que le modèle de Markov et le modèle fractal ainsi que les approches basées sur la transformée, telles que la transformée de Gabor et la transformée en ondelette ont été citées dans des études récentes (Kaur, Nazir, et Manik, 2021 ; Manimozhi et Janakiraman, 2019). Lorsque plus d'une échelle est présente dans la texture d'une image, il peut être possible d'y percevoir différentes textures et les approches multirésolutions, par exemple la transformée en ondelette pyramidale, ont obtenu de bons succès pour l'extraction des caractéristiques de ce type

d'image (Yang, Hou et Huang, 2004). L'étape de classification permet de regrouper les caractéristiques similaires. Plusieurs approches permettent de réaliser cette étape : la classification par réseau neuronal, la classification traditionnelle de Bayes ou encore un algorithme de clustering. Deux études publiées en 2008 et 2009 ont proposé des approches utilisant des méthodes traditionnelles de traitement de l'image pour faire une classification des marques d'outils. Les auteurs abordent deux problèmes principaux : comment représenter une marque d'outil et comment extraire des caractéristiques utiles de ces représentations ? La première étude explore une nouvelle approche d'analyse fractale étendue en utilisant les dimensions fractales directionnelles multiéchelles *(directional multi-scale fractal dimensions)* avec 192 images de marques d'outils appartenant à 12 classes. La seconde étude propose d'utiliser des caractéristiques d'azimut de la ligne, d'angle et de quantité de lignes, extraites du motif de la structure morphologique de la marque. Cette structure morphologique est obtenue en appliquant un opérateur Laplacien gaussien (LoG) suivi d'une transformée de Hough, sur une image en tons de gris d'une image de marque d'outil striée. Pour les expérimentations, 240 images appartenant à quatre types de structures ont été utilisées. Dans les deux études, un algorithme basé sur le clustering et la distance bayésienne est utilisé pour effectuer une classification supervisée. Les résultats de la première étude soutiennent l'efficacité des caractéristiques fractales pour la classification des marques d'outils. Les résultats de la seconde étude montrent que les caractéristiques de structure morphologique représentent adéquatement les marques d'outils, mais l'application de la technique est limitée aux stries en lignes droites et parallèles et la texture de fond de la marque d'outil n'est pas considérée (Yang *et al.*, 2008 ; Yang et Mou, 2009).

Dans les approches supervisées, des exemples de défauts étiquetés sont présentés à l'algorithme lors de l'entraînement du modèle. Puisque le processus de sélection et d'étiquetage des échantillons est une étape complexe et propice aux erreurs, certains auteurs considèrent que cette approche présente des lacunes importantes pour l'identification des défauts de surface. Ces auteurs estiment qu'une approche non supervisée serait mieux adaptée pour ce type de problème (Tomczak *et al.*, 2007).

Déjà en 2002, une étude avait proposé d'utiliser des méthodes non supervisées de clustering afin d'identifier des rayures sur des plaquettes de semi-conducteurs dans un contexte d'inspection automatisée sur des lots de production. Les auteurs ont incorporé l'indice de Calinsky Harabasz à leur méthode, un ratio entre les variances intra et intergroupes, afin de déterminer le nombre optimal de groupes à utiliser. Par la suite, une analyse de la transformée de Hough permet de classifier les défauts groupés afin d'identifier les rayures. L'article mentionne avoir évalué différents algorithmes de clustering, mais ne présente que les résultats obtenus par les algorithmes de clustering hiérarchique agglomératif (clustering par liaison complète, par liaison simple, par lien moyen, par centroïdes, à variance minimale de Ward). Deux ensembles de données ont été utilisés pour les tests : un premier ensemble de 15 plaquettes, créé artificiellement avec différents niveaux de bruit aléatoire et représentant des rayures verticales, diagonales ainsi que des défauts (non spécifiés) sur le contour, centraux et annulaires; le second ensemble provient d'images réelles de plaquettes. Les résultats supportent que l'algorithme de clustering à lien unique *(single link clustering)* offre la meilleure performance pour la détection des rayures en utilisant la transformée de Hough, mais que les résultats varient selon le type de défaut. Aussi, les auteurs ont remarqué que la transformée de Hough présente des difficultés pour la reconnaissance des défauts de contour, centraux et annulaires (Kundu, White et Mastrangelo, 2002).

En 2003, une étude a proposé une approche de clustering basée sur le chemin *(path-based)*, afin de segmenter une image selon les différentes textures qu'elle contient. L'algorithme forme les groupes en considérant l'homogénéité locale plutôt que la similarité globale, ce qui lui permet de regrouper des objets qui sont connectés par une séquence d'objets intermédiaires. Les objets considérés sont des edgels, c'est-à-dire des éléments de bord possédant une position et une direction. Afin de déterminer les edgels de l'image, un détecteur de contours (tel que Canny) est d'abord utilisé afin de trouver les pixels de bords et la direction de leur gradient. Les pixels de bords d'un voisinage local possédant des directions de gradient similaires sont utilisés pour calculer les edgels, positionnés au centre et possédant la valeur moyenne de la direction des estimés locaux. L'étape suivante consiste à éliminer les edgels bruités, qui sont nuisibles à la reconnaissance visuelle de l'image.

L'algorithme de clustering basé sur le chemin utilise ensuite les propriétés de cocircularité, de proximité et de régularité de la courbe des edgels comme mesures de similarité afin d'établir les groupes. Les éléments dont l'éloignement de tout groupe dépasse un seuil sont considérés comme étant des aberrations et sont éliminés. L'ensemble d'images publique *« Corel Image Database »* a été utilisé pour les expérimentations et les auteurs précisent que la méthode est applicable à un grand nombre de problèmes de clustering non supervisé (Fischer et Buhmann, 2003).

Une étude menée en 2004 a proposé d'utiliser un algorithme de clustering en arbre-kd *(kd-tree)* afin de segmenter les textures dans une image. Les caractéristiques de texture sont d'abord extraites à l'aide de la transformée en ondelette *(wavelet transform)* puis lissées selon un algorithme basé sur la partition en quart *(quarter partition)*, avant d'être fournies à un algorithme de clustering. Les résultats de simulation supportent que les performances de la méthode surpassent celles de l'algorithme de clustering K-moyens *(K-means)*, une approche populaire utilisée pour la classification de caractéristiques (Yang, Hou et Huang, 2004).

Dans deux études complémentaires publiées en 2006 et 2007, les auteurs ont proposé une méthode pour la détection des défauts de texture destinée à être intégrée dans un système d'inspection visuelle automatique. La technique débute avec une division de l'image d'entrée en zones non chevauchantes pour chacune desquelles une analyse statistique détermine les caractéristiques. Deux approches sont évaluées : l'analyse en composantes principales (PCA) et la décomposition en valeurs singulières (SVD). Finalement, l'algorithme de clustering C-moyens flou (FCM) permet de séparer les zones défectueuses de celles sans défauts. Les images de textures naturelles utilisées dans les expérimentations proviennent de l'ensemble de données publique *« Image After »*, qui propose des échantillons avec et sans défauts. Les résultats démontrent la capacité de la méthode à détecter efficacement les défauts pour différents types de matériaux, notamment le bois, la roche et les surfaces peintes. Cependant, les auteurs suggèrent des études supplémentaires pour d'autres types de matériaux. De plus, les résultats supportent que la méthode proposée surpasse en efficacité et en précision deux

classificateurs de texture supervisés (KNN et un classificateur à fonction discriminante), que les auteurs ont testés sur les mêmes données (Tomczak, Mosorov et Sankowski, 2006 ; Tomczak *et al.*, 2007).

En 2019, une étude a tenté de détecter des défauts sur des surfaces texturées. Pour extraire les caractéristiques de la texture, les auteurs ont testé deux options : la matrice de cooccurrence de niveau de gris (GLCM) et la matrice de longueur de course de niveau de gris (GLRLM). Ensuite, les caractéristiques extraites sont classifiées par une machine à vecteur de support améliorée par la combinaison d'un noyau linéaire et d'un noyau quadratique, dans le but de favoriser la classification. Finalement, un algorithme de clustering de C-moyens flou (FCM), modifié par l'ajout d'un mécanisme de sélection des centroïdes, permet de segmenter les caractéristiques classées afin d'identifier les défauts sur la surface. L'ensemble de données publique *« Tilda texture dataset »* a été utilisé pour faire les expérimentations et les auteurs considèrent que les résultats obtenus dépassent ceux de l'état de l'art, notamment avec une sensibilité de 98,92%, une spécificité de 99,59% et une précision globale de 99,59% (Manimozhi et Janakiraman, 2019).

Les auteurs d'une étude de 2020 ont aussi utilisé une méthode de clustering afin de détecter des rayures sur des images de plaques de céramique piézoélectrique en tons de gris, présentant une texture irrégulière avec des valeurs de gris aléatoires. Dans cette étude, les auteurs utilisent d'abord l'algorithme de clustering C-moyens flou FCM *(Fuzzy C-means)* jumelé à des méthodes d'interpolation pour diviser les pixels de l'image selon leur appartenance au groupe de l'avant-plan et obtenir une valeur du degré d'appartenance pour chaque pixel. Ensuite, l'image grise originale est améliorée en la multipliant avec ces valeurs d'appartenance, puis en effectuant des opérations de normalisation et de binarisation. Dans la seconde partie de la méthode, les caractéristiques morphologiques de l'image permettent de segmenter les rayures et filtrer celles dont le nombre de pixels est inférieur à un seuil établi. Puis, les informations de localisation entre les régions et de direction du gradient des régions permettent de fusionner les éléments appartenant à la même rayure et de déterminer si une zone de segmentation est rayée ou non. Lors des expérimentations, les auteurs ont observé

une invariance à la rotation et à la translation en analysant les résultats de la même pièce sous vingt angles et positions différentes. Les résultats obtenus soutiennent que la méthode permet de détecter avec précision les rayures sur une surface de céramique dont la texture est aléatoire. Par contre, les travaux sont limités à l'étude d'une trentaine d'échantillons, et seulement la détection des rayures est considérée : pour les travaux futurs, les auteurs proposent d'étendre leurs études à d'autres types de marques (Tan *et al.*, 2020).

Dans une étude menée en 2021, les auteurs ont tenté de regrouper des circuits intégrés défectueux sur des cartes de plaquettes de silicium *(wafer maps)* afin d'implémenter un système d'inspection automatique. Ils ont travaillé à partir de 6 722 cartes et leur méthodologie comprend trois phases : l'extraction des caractéristiques à partir d'images filtrées des cartes, le regroupement des caractéristiques selon leur similarité et une étape d'optimisation durant laquelle les groupes possédant des noyaux similaires sont fusionnés, réduisant la similarité interclusters et augmentant ainsi la performance. Les auteurs demeurent discrets sur l'algorithme de clustering utilisé, mais ils définissent une métrique de similarité alternative nommée min-max par élément *(element-wise min-max)*. Les résultats suggèrent que la performance de leur méthode dépasse celle d'OPTICS, une méthode de clustering bien connue (Li *et al.*, 2021).

Quelques études ont utilisé des méthodes de clustering non supervisé afin de détecter des défauts dans des images texturées en treillis *(lattice)*, comme des tissus ou des papiers peints. Les travaux sont intéressants, mais les méthodes ne sont pas nécessairement applicables à la recherche de défauts sur des surfaces texturées générales puisqu'ils se basent sur le motif de base de la surface pour faire les regroupements et identifier les dissemblances (Ngan, Pang et Yung, 2007 ; Ngan, Pang et Yung, 2010). Dans le cas de la surface d'une douille de cartouche, la texture est plutôt aléatoire, sans motifs de texture se répétant dans l'image.

De façon similaire, une étude publiée en 2013 propose une solution basée sur l'algorithme de décalage moyen *(Mean-Shift)* afin de détecter, regrouper et segmenter des motifs répétitifs dans une image. Les images utilisées dans les expérimentations proviennent de la base

publique *« PSU-Near Regular Texture Database »* et possèdent des motifs répétés distinctifs (par exemple des visages dans un groupe, des pétales sur une fleur ou des motifs dans un treillis). Il serait intéressant de voir si la méthode peut s'appliquer à la détection de marques répétitives plus subtiles. Par contre, l'étude n'identifie qu'un seul motif répété par image (Cai et Baciu, 2013).

L'analyse des textures, quant à elle, permet de décrire et de quantifier une texture perçue par des termes tels que rugueux, lisse, irrégulier ou hachuré. Une étude de 2022 souligne l'arrivée récente de l'analyse des textures dans le domaine de l'imagerie médicale et la décrit comme une avancée dans l'évaluation des maladies (Tang *et al.*, 2022). Dans cette étude, les auteurs ont utilisé le modèle binaire local (LBP) combiné à la matrice de cooccurrence des niveaux de gris (GLCM) pour extraire et analyser les caractéristiques dans des images échographiques du muscle droit du fémur. Les valeurs de contraste (mesure la variation des niveaux de gris), d'entropie (mesure le désordre dans la texture) et d'homogénéité (mesure le changement dans différentes parties de la texture) sont retenues pour distinguer des changements dans la texture musculaire de la sarcopénie. L'étude se concentre sur la robustesse de la méthode face aux changements de luminosité dans les images et les résultats supportent que la méthode soit stable pour l'extraction de caractéristiques de la texture musculaire (Tang *et al.*, 2022).

L'apprentissage profond ne demeure pas à l'écart du clustering et une étude menée en 2020 s'est tournée vers les réseaux neuronaux pour identifier des motifs de textures dans des images médicales de résonance magnétique, afin de déterminer le stade de la maladie de stéatose hépatique. Pour ce faire, ils ont utilisé un réseau de clustering profond (DCN), qui peut prendre en charge le codage, la réduction de la dimensionnalité et le regroupement des caractéristiques de l'image en combinant des couches d'autoencodeurs et de décodeurs. Les groupes ainsi formés servent ensuite de caractéristiques pour le classificateur de forêt aléatoire. L'étude a utilisé 70 images IRM et les résultats supportent que cette approche permette d'identifier les différents stades de la maladie et de capturer sa progression même si aucune information a priori n'est disponible (Perkonigg *et al.*, 2020).

En terminant, une étude en balistique menée en 2001 avait aussi utilisé une technique de clustering afin d'identifier le type d'arme à feu utilisé pour tirer une balle et sa cartouche. Mais dans leur cas, les regroupements sont basés sur des paramètres mesurables des spécimens balistiques (tels que le calibre, la forme du percuteur, la longueur de la douille) et non sur des images. L'étude mentionne avoir aussi utilisé les informations de l'empreinte du percuteur, de la marque de l'éjecteur et de la marque de l'extracteur, mais ne spécifie pas comment ces marques ont été mesurées. Les caractéristiques d'un spécimen sont ensuite représentées sur un graphique de n dimensions, puis elles sont associées à un groupe connu; mais aucune précision n'est fournie sur les techniques utilisées. Les auteurs concluent que leur méthode est prometteuse pour l'identification du type d'arme à feu, mais ils soulignent l'importance d'obtenir des mesures précises pour les caractéristiques de classe. De plus, l'étude s'est limitée à un nombre restreint de spécimens : seulement quatorze cartouches et projectiles ont été évalués (Smith, 2001).

CHAPITRE 2

OBJECTIFS

2.1 Objectifs

2.1.1 Objectif principal

L'objectif principal de cette thèse est de déterminer une méthode d'apprentissage machine permettant d'identifier des marques microscopiques présentes sur des images de douilles de cartouche.

2.1.2 Objectifs spécifiques

1) Classifier les marques microscopiques présentes sur les douilles de cartouche à l'aide de méthodes d'apprentissage profond supervisé
2) Évaluer des regroupements d'images de douilles de cartouche, obtenus à l'aide de méthodes d'apprentissage non supervisé
3) Étudier l'accord interobservateur entre des observateurs humains et les méthodes d'apprentissage profond supervisé

2.1.3 Contributions

À ce jour, peu de travaux se sont intéressés aux méthodes basées sur l'apprentissage profond afin de classifier des images de douilles de cartouches. Les marques de classe sur les douilles de cartouche sont de petites tailles et s'apparentent à une texture de surface. Les différents types de marques sont très similaires les uns aux autres et ils sont difficilement discernables à l'observation. Quelques études ont cherché à identifier des marques variées et des défauts sur des surfaces texturées, mais sur d'autres types de surfaces. Par exemple, des marques de coupures sur des os ou des rayures sur des plaquettes de silicium. Quelques études ont évalué différentes méthodes supervisées visant à comparer et à apparier deux douilles de cartouche,

mais pas dans le but d'en identifier les marques de classe. Les méthodes d'apprentissage supervisé, autant avec des méthodes traditionnelles de vision par ordinateur accompagnées d'apprentissage machine qu'avec des méthodes plus récentes impliquant des réseaux neuronaux, ont obtenu de bons résultats pour la classification de marques avec un haut niveau de similitude.

Ainsi, la **première contribution** explorera différentes méthodes d'apprentissage supervisé afin de classifier les marques microscopiques présentes sur les douilles de cartouche. Pour cette tâche, la principale difficulté sera de trouver une technique ou une architecture qui parviendra à discriminer les différences subtiles; les solutions permettant une classification à grain fin seraient sans doute mieux adaptées pour ce problème. En ce qui concerne les méthodes traditionnelles, des techniques multiéchelles pour l'extraction des caractéristiques, accompagnées d'un classificateur simple, semblent avoir eu du succès dans des études dont les images présentaient des textures de surface. Du côté de l'apprentissage profond, plusieurs choix d'architectures sont possibles afin d'effectuer des classifications multiclasses et multiétiquettes. En outre, il serait intéressant de comparer les performances obtenues avec une approche CNN à celles obtenues avec une approche ViT. Dans des travaux préliminaires publiés en 2023, nous avons exploré l'apprentissage supervisé avec la segmentation de la région d'intérêt sur des images de douilles de cartouche (Le Bouthillier *et al.*, 2023).

La revue de littérature démontre aussi que peu d'études ont utilisé l'apprentissage profond afin de détecter des marques sur des surfaces texturées par des méthodes de clustering non supervisées. Malgré tout, plusieurs études semblent avoir eu du succès avec des méthodes de clustering traditionnelles et il serait intéressant d'évaluer leur efficacité avec la tâche de regroupement des marques de classe sur les douilles de cartouche, ce qui constitue la **deuxième contribution**. De plus, une méthode de clustering utilisant un réseau neuronal pourrait possiblement améliorer les performances obtenues. À cet effet, il serait intéressant d'évaluer une méthode de clustering profond et de comparer les résultats avec ceux obtenus par les algorithmes de clustering traditionnels.

La **troisième et dernière contribution** a été élaborée à la suite des expérimentations des deux premiers objectifs, au cours desquelles des irrégularités ont été observées dans les étiquettes de la vérité de terrain. Dans un article proposant d'utiliser l'apprentissage profond afin d'évaluer le stade de Risser dans des cas de scoliose idiopathique chez l'adolescent (Kaddioui *et al.*, 2020), les auteurs ont utilisé des métriques interobservateurs afin de comparer les résultats de leur modèle avec les évaluations des experts humains. Nous nous sommes inspirés de ces travaux afin d'ajouter des métriques interobservateurs, permettant de mieux comprendre la difficulté de la tâche d'annotation des marques microscopiques sur les douilles de cartouche, et par la même occasion, de mieux situer les performances des algorithmes vis-à-vis des experts humains.

Le diagramme de la Figure 2.1 illustre ces trois contributions, ainsi que les liens qu'elles ont entre elles. En outre, les modèles entraînés lors du premier objectif seront utilisés pour extraire des caractéristiques descriptives des images, qui seront utilisées par les algorithmes de clustering du second objectif. Un ensemble de vérité de terrain vérifié sera construit au cours du dernier objectif, et il sera utilisé afin d'entraîner de nouveaux modèles de prédiction, dont les résultats seront comparés à ceux obtenus au cours du premier objectif.

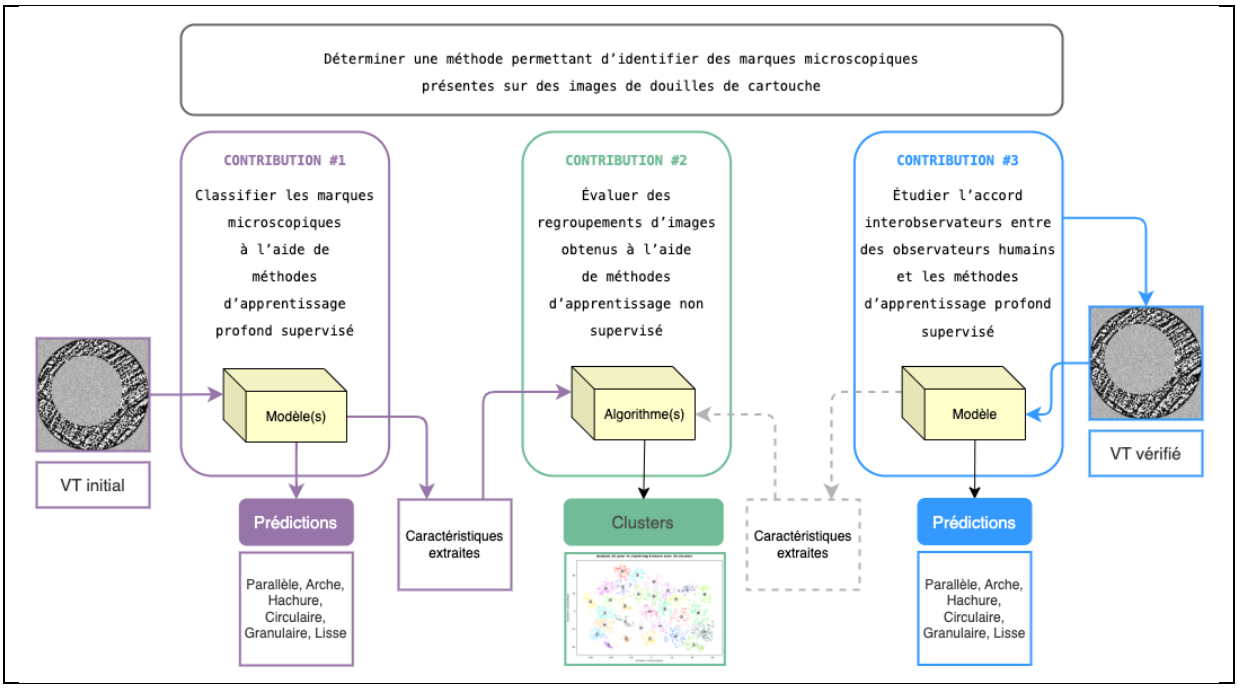


Figure 2.1 Contributions de la thèse : la contribution #1 (classifier les marques microscopiques), la contribution #2 (évaluer des regroupements d'images) et la contribution #3 (étudier l'accord interobservateur)

CHAPITRE 3

DESCRIPTION DES DONNÉES

Ce chapitre présente les données qui ont été utilisées lors des expérimentations. Ces données proviennent du département de recherche de l'entreprise ***LeadsOnline***, anciennement ***Ultra Electronics Forensic Technology***, une entreprise spécialisée dans le développement d'équipements destinés à l'identification des pièces à conviction balistiques. Ses systèmes automatisent l'acquisition d'images de balles et de douilles issues du tir d'une arme à feu, puis proposent des jumelages avec d'autres pièces à conviction en se basant sur un score de similitude calculé par un algorithme privé. Les examinateurs peuvent ensuite analyser ces résultats sur des stations de visualisation afin de confirmer ou de rejeter les appariements proposés. Le processus d'acquisition gère le positionnement de l'échantillon balistique, à l'exception de son orientation, qui doit être ajustée manuellement par le technicien, ainsi que la mise au point de l'image, et la délimitation de la région d'intérêt (ROI). Deux types d'images sont produites : des images de luminance en 2D, réalisées avec une source lumineuse annulaire ou rasante; et des images topographiques en 3D, présentant une résolution verticale de l'ordre du micron (UEFT, [s d] ; LO, [s d]).

3.1 Description des images

Les images de la base de données présentent différentes marques microscopiques situées sur des images de la marque de la face de la culasse (BF) d'une douille de cartouche (Figure 3.1). Six types de marques sont possibles: parallèles, arches, hachures, circulaires, granulaires ou lisses, et plus d'un type de marque peut être visible sur une même douille. Une septième catégorie inclut les échantillons exhibant des marques d'un type inconnu. La Figure 3.2 présente quelques exemples.

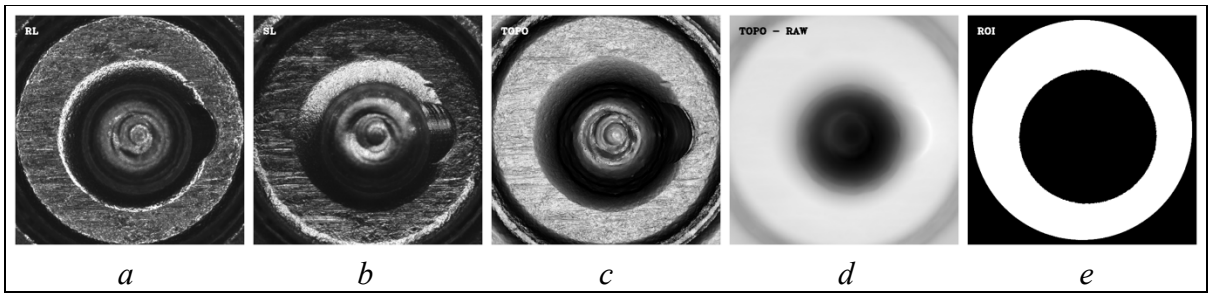


Figure 3.1 Images BF : photographiées avec un éclairage annulaire (a) et un éclairage latéral (b), une représentation 2D de la topographie (c),la topographie 3D (d) et le masque binaire de la région d'intérêt (e)

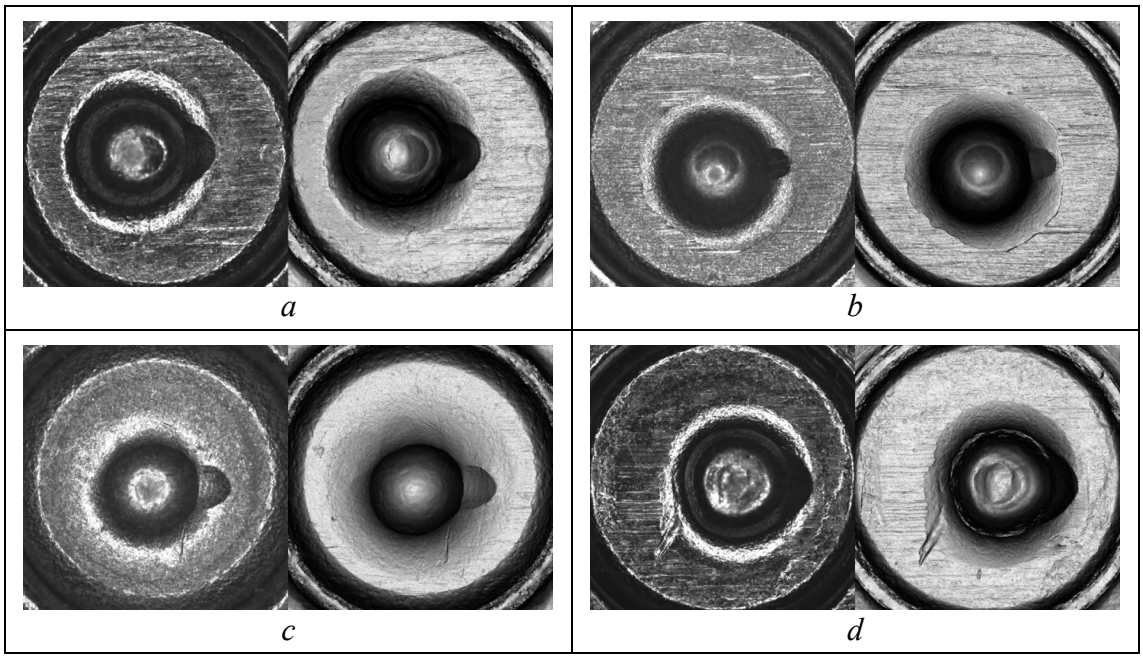


Figure 3.2 Marques sur des douilles de cartouche, capturée avec une source lumineuse annulaire (gauche), topographie 3D (droite); hachures + granulaires (a), parallèles + hachures + granulaires (b), parallèles + hachures (c), parallèles + granulaires (d)

3.2 Description des jeux de données

Deux jeux de données étiquetés ont été utilisés pour les expérimentations. Le premier ensemble est composé de 9876 échantillons provenant des acquisitions d'un appareil BrassTRAX de troisième génération (HW3) et dont les images sont de 960 X 960 pixels, avec une résolution de 4.8 microns par pixel. Le second ensemble est composé de 3609

échantillons provenant des acquisitions d'un appareil BrassTRAX de la quatrième génération (HW4) et dont la taille varie pour chaque image, se situant autour de 1400 X 1400 pixels, avec une résolution de 3.3 microns par pixel. Les deux ensembles d'images couvrent le même champ de vue, c'est-à-dire la même aire de la douille. Nous disposons aussi d'un ensemble de 7 402 échantillons non étiquetés, provenant de deux autres appareils BrassTRAX (HW4).

Pour chacune des douilles, les ensembles contiennent des images 2D et la topographie 3D de la marque de la face de la culasse BF, ainsi que les coordonnées de deux cercles délimitant cette zone, à partir desquels il est possible de générer un masque binaire de la région d'intérêt. Pour les ensembles étiquetés, un expert a identifié les catégories de marques de classe pour chaque échantillon. Ces annotations correspondant à la vérité de terrain *(ground truth),* sont conservées dans un fichier CSV. Il n'y a pas de masques ou de coordonnées indiquant la localisation des marques de classe sur les images, seulement les catégories sont disponibles.

Les ensembles de données contiennent plusieurs paires provenant de l'acquisition de douilles de cartouches sœurs; c'est-à-dire de cartouches distinctes ayant été tirées par la même arme à feu. Les cartouches sœurs ont généralement un niveau de similarité plus élevé que les cartouches tirées par des armes à feu différentes. Les ensembles de données sont ainsi créés afin de pouvoir tester les performances des algorithmes d'appariement. Après une inspection visuelle des images, nous avons choisi de conserver ces cartouches sœurs, car nous jugeons que les images sont suffisamment différentes les unes des autres pour ne pas risquer de fausser la précision des résultats obtenus, par exemple si une image de douille se retrouve aléatoirement dans un ensemble d'entraînement alors que l'image de sa sœur se retrouve dans l'ensemble de tests.

3.3 Analyses exploratoires

3.3.1 Données étiquetées du BrassTRAX HW3

Lors de l'inspection initiale, nous avons remarqué que 54 échantillons étaient nommés avec les termes « ERROR », « ERRO », ou « ERR ». Ces échantillons sont des doublons comportant différentes erreurs survenues lors de la prise des images, et nous les avons retirés de l'ensemble. L'analyse exploratoire qui suit a été obtenue à partir d'une liste de 9822 échantillons. De plus, nous avons remarqué que 16 images étaient manquantes pour les topographies 3D. L'ensemble d'images topographiques se trouve ainsi réduit à 9806 échantillons. Le graphique de la Figure 3.3 (a) détaille le nombre d'échantillons contenant des marques appartenant à chaque catégorie. Puisque le problème est multiétiquette, il est possible que le même échantillon soit calculé dans plus d'une catégorie. C'est pourquoi le nombre total d'échantillons représentés dans le graphique est supérieur au nombre d'échantillons réels. Dans ce graphique. On remarque que la catégorie granulaire est beaucoup plus présente que les autres catégories. La Figure 3.3 (b) nous permet de visualiser la distribution des calibres pour les douilles. On remarque que les calibres 9LUG, 45AU et 380A sont prédominants.

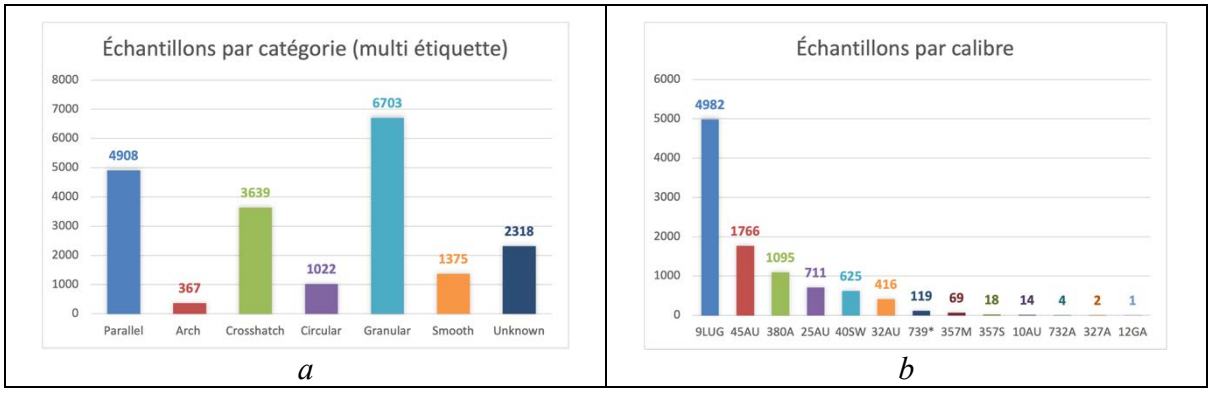


Figure 3.3 BrassTRAX HW3, multiétiquette : nombre d'échantillons par catégorie (a), nombre d'échantillons par calibre (b)

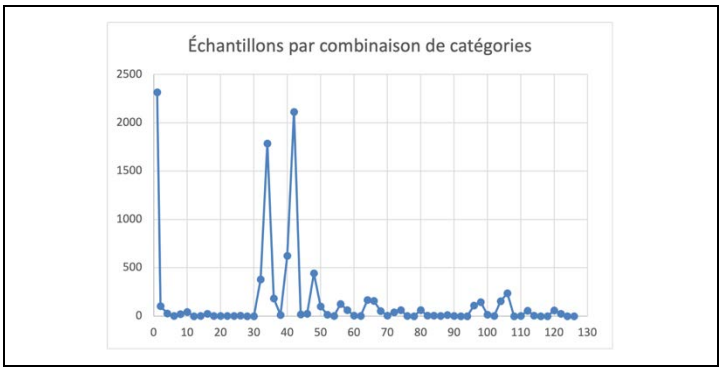


Figure 3.4 BrassTRAX HW3 : nombre d'échantillons par combinaison de catégories

Avec six catégories de marques plus une catégorie inconnue, nous avons la possibilité de retrouver 64 combinaisons de catégories. Chaque combinaison de catégories est représentée par un chiffre unique, dont la valeur permet de reconstituer le contenu. Par exemple, les catégories des sept types de marques sont représentées par 1 : inconnue, 2 : parallèle, 4 : arche, 8 : hachure, 16 : circulaire, 32 : granulaire et 64 : lisse. Ainsi la catégorie 34 représente la combinaison de marques parallèle et granulaire (2 + 32); et la catégorie 82 représente la combinaison de marques parallèle, circulaire et lisse (2 + 16 + 64). La Figure 3.4 représente la distribution du nombre d'échantillons par combinaisons de catégories. Nous pouvons observer que les combinaisons 34 (parallèle et granulaire) et 42 (parallèle, hachure et granulaire) sont celles qui possèdent le plus d'échantillons (1786 et 2113, respectivement). Les combinaisons 32 (parallèle), 40 (hachure et granulaire) et 48 (circulaire et granulaire) sont moins bien représentées, mais dépassent le reste des catégories avec 381, 623 et 444 échantillons, respectivement. Pour le reste des combinaisons possibles, moins d'échantillons sont disponibles. Dix combinaisons de catégories possèdent entre 100 et 250 échantillons, dix combinaisons possèdent entre 25 et 99 échantillons, et 27 combinaisons possèdent moins de 25 échantillons. 11 combinaisons de catégories n'ont aucun échantillon permettant de les représenter (parmi celles-ci se retrouve la combinaison irréaliste contenant les six types de marques). Ainsi, nous remarquons que les données sont déséquilibrées à deux niveaux. D'abord, l'ensemble contient beaucoup plus d'échantillons exhibant les marques parallèle, granulaire et hachure. Ensuite, certaines combinaisons de catégories sont plus présentes que d'autres.

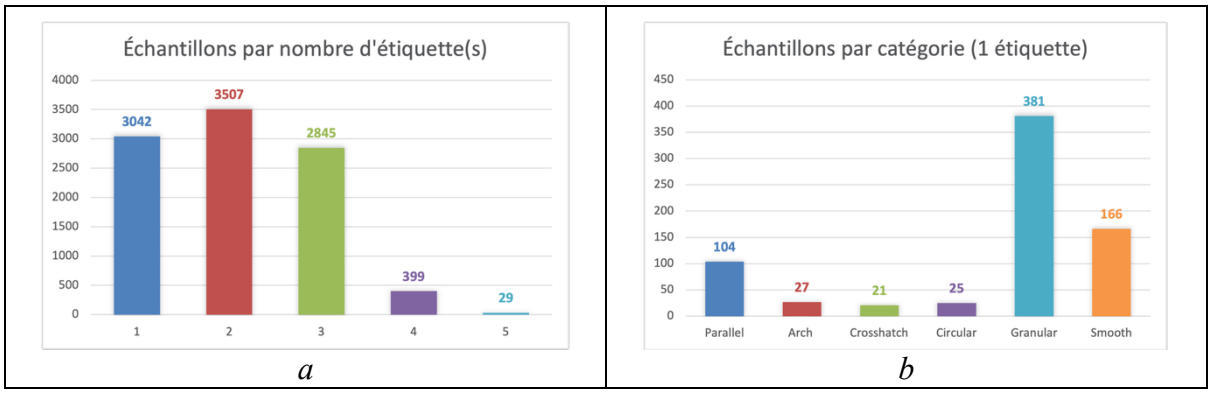


Figure 3.5 BrassTRAX HW3 : nombre d'échantillons par nombre d'étiquettes (a), nombre d'échantillons par catégorie (échantillons avec une seule étiquette) (b)

La Figure 3.5 (a) représente le nombre d'échantillons par nombre d'étiquettes. On peut voir que, parmi les 9822 échantillons, 3042 ne comportent qu'une seule étiquette, 3507 en comportent deux, 2845 en ont trois alors que peu d'échantillons possèdent quatre ou cinq étiquettes (399 et 29, respectivement). Aucun échantillon ne comporte toutes les étiquettes (6). Les échantillons appartenant à la 7e catégorie (inconnu) ne sont jamais multiétiquettes et sont donc comptabilisés avec les échantillons comportant une seule étiquette. Parmi les 3042 échantillons n'ayant qu'une étiquette, 2318 appartiennent à la catégorie « inconnu ». En reprenant l'analyse sur les 724 échantillons restants, nous obtenons le graphique représenté dans la Figure 3.5 (b), qui présente le nombre d'échantillons par catégorie de marques. Nous pouvons observer que les marques granulaires sont majoritaires, suivies par les marques lisses et parallèles.

3.3.2 Données étiquetées du BrassTRAX HW4

Le graphique de la Figure 3.6 (a) détaille le nombre d'échantillons contenant des marques appartenant à chaque catégorie. Tout comme avec les données provenant du BrassTRAX HW3, nous remarquons que les catégories granulaire et inconnu sont beaucoup plus présentes que les autres catégories. La Figure 3.6 (b) nous indique que les calibres 9LUG, 45AU et 380A sont toujours prédominants.

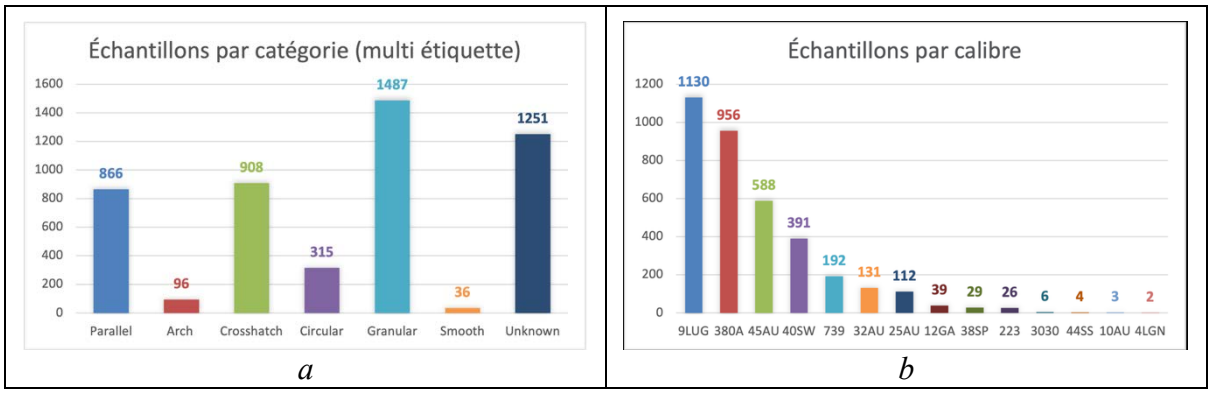


Figure 3.6 BrassTRAX HW4, multiétiquette : nombre d'échantillons par catégorie (a), nombre d'échantillons par calibre (b)

Sur le graphique de la Figure 3.7, nous pouvons observer que les combinaisons 2 (parallèle), 8 (hachure), 32 (granulaire), 34 (parallèle et granulaire) et 40 (hachure et granulaire) sont celles qui possèdent le plus d'échantillons (entre 276 et 426). Pour le reste des combinaisons, nous observons six combinaisons de catégories contenant entre 25 et 250 échantillons et dix-huit combinaisons possédant moins de 25 échantillons. Trente-quatre combinaisons de catégories ne possèdent aucun échantillon.

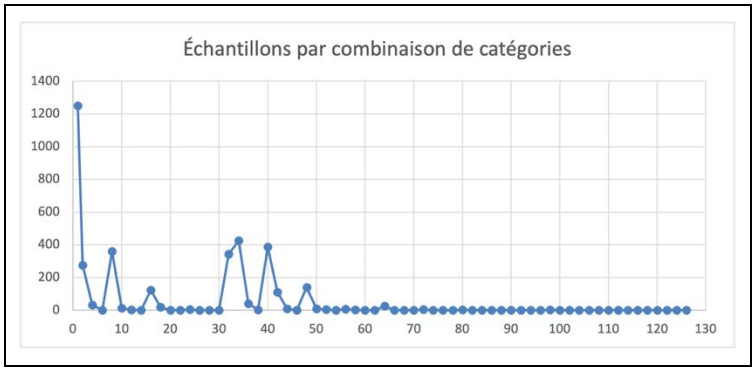


Figure 3.7 BrassTRAX HW4 : nombre d'échantillons par combinaison de catégories

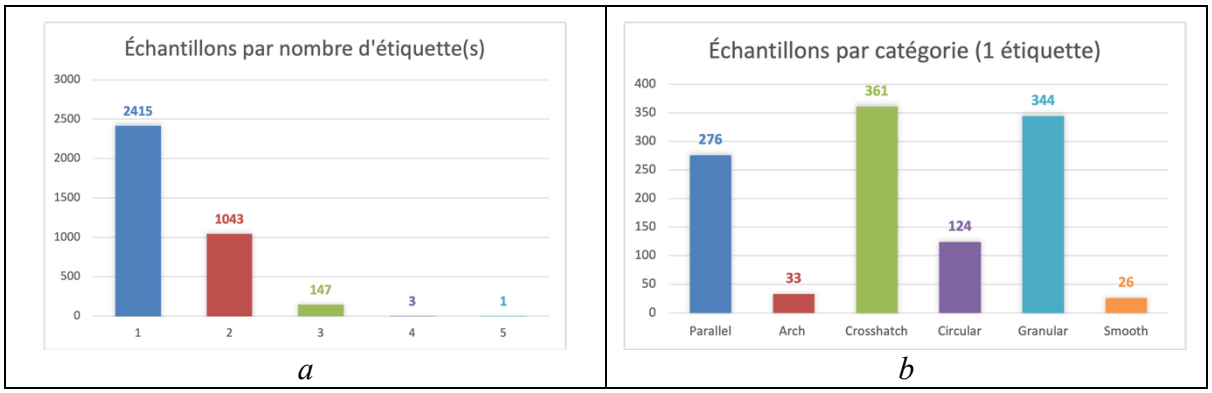


Figure 3.8 BrassTRAX HW4 : nombre d'échantillons par nombre d'étiquettes (a), nombre d'échantillons par catégorie (échantillons avec une seule étiquette) (b)

Sur le graphique de la Figure 3.8 (a), nous remarquons que, parmi les 3609 échantillons disponibles, 2415 ne comportent qu'une seule étiquette, 1043 en comportent deux, alors que seulement 151 échantillons comportent plus de trois étiquettes. Nous considérons cela comme étant positif, car les échantillons possédant plus de trois étiquettes devraient être des cas plus rares. Finalement, le graphique de la Figure 3.8 (b) présente la distribution des catégories pour les échantillons avec une seule étiquette (excluant les 1251 échantillons de catégorie inconnue). Nous pouvons observer que la marque hachure est majoritaire, suivie de près par les marques granulaire et parallèle.

3.4 Prétraitements des images topographiques

Les topographies 3D doivent être traitées afin de présenter une image 2D compréhensible à l'affichage, et avant d'être fournies en entrée à un réseau de neurones. Cette section présente les étapes nécessaires à ce traitement. Les autres prétraitements, spécifiques aux expérimentations, seront abordés dans les chapitres traitant des contributions.

Tout d'abord, les valeurs des pixels des échantillons 3D sont enregistrées en format court non signé (unsigned short), un entier dont la valeur se situe entre 0 et 65 535. De plus, chaque échantillon possède un facteur d'échelle associé à l'image topographique, sauvegardé séparément dans la base de données. La première étape, lorsqu'une image topographique est chargée, est de la transformer dans une échelle [0, 1] en divisant chaque pixel par 65 535 et

en multipliant le résultat par 1000 et par la valeur de l'échelle propre à l'échantillon (65 535 * 1000 * facteur d'échelle)

Nous appliquons ensuite un filtre lissant sur l'image topographique, afin de faire ressortir les grosses structures de l'image. Puis, cette image filtrée est soustraite de l'image originale, afin de mettre les détails de l'image en évidence. Après avoir expérimenté avec différentes techniques de filtres et plusieurs valeurs de paramètres (dont les détails ne seront pas présentés dans cette thèse), nous avons sélectionné un filtre gaussien provenant de la librairie OpenCV, avec un noyau (kernel) de 21 et un sigma de 6. La dernière étape de ce traitement est de transformer l'image filtrée en entier non signé (UINT8).

CHAPITRE 4

CONTRIBUTION #1 – APPRENTISSAGE PROFOND SUPERVISÉ POUR LA CLASSIFICATION DE MARQUES MICROSCOPIQUES

4.1 Introduction

Ce chapitre discute du premier objectif : « Classifier les marques microscopiques présentes sur les douilles de cartouche à l'aide de méthodes d'apprentissage profond supervisé ». Il débute avec une présentation de la méthodologie; incluant le choix des algorithmes et des métriques, la préparation des données, ainsi que le détail de l'implémentation. Le chapitre se poursuit avec une présentation des résultats pour les différentes expérimentations, accompagnée d'observations.

4.2 Méthodes

La classification des marques sur les douilles de cartouche est un problème multiclasse, c'est-à-dire qu'il existe plus de deux catégories de marques possibles; et multiétiquette, c'est-à-dire qu'un échantillon peut appartenir à plus d'une catégorie. Nous avons divisé les expérimentations en trois parties : classification multiclasse (pour les échantillons possédant une seule étiquette), classification multiétiquette (pour tous les échantillons) et classification binaire (un classificateur pour chaque catégorie).

Pour ce chapitre, nous avons décidé d'ignorer les échantillons étiquetés « inconnu », étant donné que les échantillons de cette catégorie sont particuliers et qu'ils ajoutent de la complexité au problème. En outre, ces échantillons ne montrent pas de marques descriptives permettant de les associer à une des catégories connues, et par conséquent, il n'existe pas de marques descriptives spécifiques appartenant à la catégorie des inconnus. Ces échantillons pourraient être considérés comme étant des valeurs aberrantes et devront possiblement être traités séparément du problème de classification.

4.2.1 Architectures pour l'analyse automatique d'images

Pour faire la classification, nous avons sélectionné deux architectures de CNN. Afin d'établir un référentiel, nous avons d'abord choisi l'architecture VGG16 qui a fait ses preuves dans de nombreuses études. VGG16 réduit la taille du filtre de convolution à 3×3, ce qui lui permet de pouvoir augmenter la profondeur du réseau. Le réseau est ainsi constitué de 16 couches, divisées en 13 couches convolutives et trois couches entièrement connectées en sortie (Simonyan et Zisserman, 2015). Nous avons aussi sélectionné l'architecture EfficientNet (Tan et Le, 2020), puisqu'elle proclame être en mesure de surpasser plusieurs CNN de l'état de l'art, et que cette affirmation est supportée par des travaux récents (Atila et al., 2021). Le réseau EfficientNet B0 a été élaboré à partir d'une architecture de réseau CNN mobile puis augmenté avec une méthode de mise à l'échelle composée, afin de créer les réseaux EfficientNet B1 à B7. Nous avons aussi sélectionné l'architecture ViT originale (Dosovitskiy *et al.*, 2021). Les mécanismes d'autoattention peuvent potentiellement aider à la détection des marques situées à différents endroits de l'image.

4.2.2 Métriques

Les courbes d'apprentissage d'entraînement et de validation pour la performance et l'optimisation ont guidé les choix des hyper paramètres lors des entraînements des différents modèles. Les métriques d'exactitude (accuracy), de précision, de rappel et la mesure F1 ont été utilisées afin d'évaluer les prédictions des classificateurs sur les ensembles de test, et les matrices de confusion ont été observées. Le coefficient de corrélation de Matthews, les courbes ROC et la mesure AUROC ont été ajoutées pour l'évaluation des modèles binaires. Finalement, la technique de visualisation GradCam a permis de produire des cartes thermographiques indiquant les zones de l'image qui sont importantes pour les modèles afin de tenter d'améliorer leur explicabilité (Selvaraju *et al.*, 2020).

4.2.3 Prétraitements sur les images originales

Avant de débuter les expérimentations, nous avons testé sommairement différentes configurations d'entrée pour les réseaux. Par exemple, l'image originale seulement, l'image originale fusionnée avec un masque de bruit gaussien aléatoire, etc. Nous avons aussi testé des modèles avec plus d'une entrée. Par exemple, deux images en entrée dont la première correspond à l'échantillon et la seconde au masque binaire. Ces tests préliminaires ont permis de sélectionner une configuration pour l'image d'entrée des modèles.

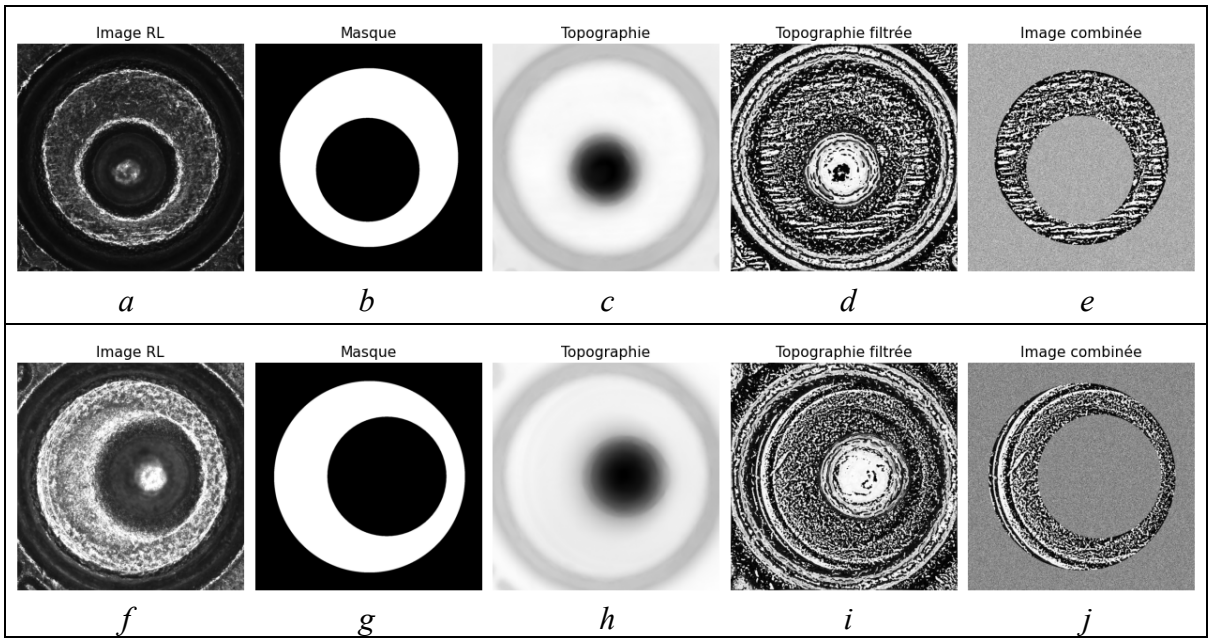


Figure 4.1 Image originale 2D (a, f), masque correspondant (b, g), topographie (c, h), topographie filtrée (d, i), topographie filtrée, sur laquelle le masque binaire a été superposé et remplacé par un bruit gaussien aléatoire (e, j)

Nous n'avons pas remarqué de différence significative parmi les différentes configurations, alors nous avons choisi d'utiliser l'image 3D filtrée (section 3.4), fusionnée avec le masque de la région d'intérêt remplacé par du bruit gaussien aléatoire (Figure 4.1 e et j), pour laquelle les résultats nous semblaient légèrement supérieurs lors de la comparaison des courbes d'apprentissage. Ces expérimentations sont un point de départ et des tests supplémentaires seraient nécessaires afin de bien évaluer ces différentes configurations pour

l'image d'entrée. La Figure 4.1 présente deux exemples de topographies transformées afin d'obtenir l'image d'entrée.

4.2.4 Préparation des données

Nous avons utilisé les outils de la librairie scikit-learn (Scikit-learn, [s d]) pour diviser les données en trois ensembles stratifiés, afin de respecter la distribution des classes : un ensemble d'entraînement comprenant 72% des échantillons, un ensemble de validation comprenant 15% des échantillons et un ensemble de test comprenant 13% des échantillons. Les redimensionnements ainsi que les prétraitements spécifiques aux différents modèles sont appliqués directement lors des entraînements ou des évaluations. Pour VGG16 et les ViT, les librairies utilisées fournissent des fonctions pour les prétraitements nécessaires. Les modèles EfficientNet ne requièrent aucun prétraitement, ces étapes sont incluses directement dans les premières couches des modèles.

Pour les expérimentations multiclasses, nous avons utilisé les échantillons comprenant une seule étiquette, excluant la catégorie « inconnu ». Au départ, nous avons entraîné et testé les modèles de CNN avec les 724 images provenant du BrassTRAX HW3 (section 3.3.1). Par la suite, nous avons reproduit ces expérimentations avec les 1164 images provenant du BrassTRAX HW4 (section 3.3.2). Pour terminer, nous avons effectué des expérimentations additionnelles avec quatre modèles de ViT sur les données provenant du BrassTRAX HW4. Ensuite, nous avons utilisé des techniques d'augmentation afin d'équilibrer les données des ensembles d'entraînement et atteindre une meilleure généralisation. Afin d'utiliser le maximum d'échantillons disponibles, nous avons augmenté individuellement chaque classe, afin d'obtenir 500 échantillons pour chacune des six catégories. Les échantillons augmentés ont été choisis aléatoirement parmi tous les échantillons de l'ensemble d'entraînement appartenant à cette classe. Par exemple, pour les images du BrassTRAX HW3, les images de la catégorie parallèle ont bénéficié d'une augmentation de 425 échantillons, alors que celles de la catégorie granulaire n'ont eu que 225 augmentations. Les opérations d'augmentation ont été choisies aléatoirement parmi des flips horizontaux, des flips verticaux et des rotations

d'un degré aléatoire. À la fin de ces opérations d'augmentation, les ensembles d'entraînement comprennent 3000 échantillons, 500 pour chacune des six catégories (parallèle, arche, hachure, circulaire, granulaire et lisse). Une fois les données divisées et augmentées, nous avons appliqué les prétraitements décrits dans la section précédente, afin d'obtenir la configuration d'entrée de l'image. Le diagramme de la Figure 4.2 illustre les étapes de préparation des données pour les expérimentations multiclasses.

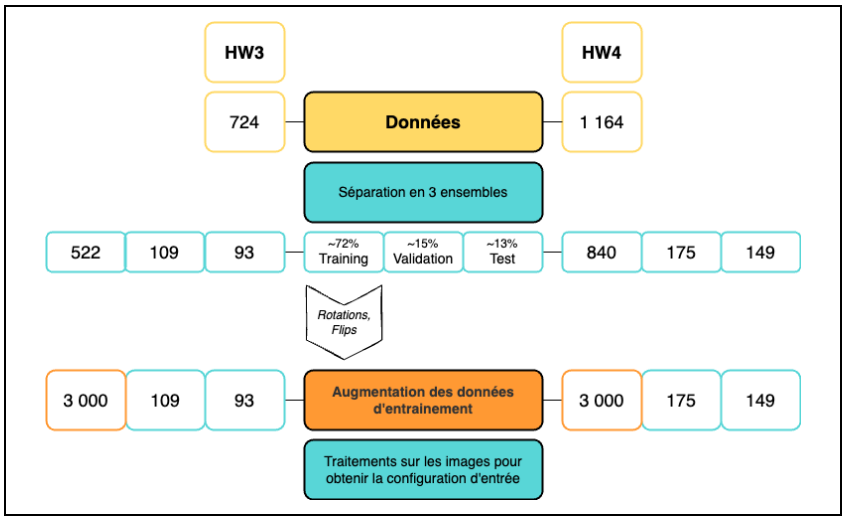


Figure 4.2 Division et augmentation des données pour les expérimentations multiclasses

Pour les expérimentations multiétiquettes, nous avons utilisé une sélection d'échantillons comprenant une ou plusieurs étiquettes provenant du BrassTRAX HW4. Tel que présenté à la Figure 3.7, plusieurs configurations d'étiquettes ont peu (ou pas du tout) d'échantillons pour les représenter. Nous avons choisi les combinaisons des catégories ayant au moins dix échantillons. Nous obtenons ainsi un ensemble de données de 2311 échantillons, composé des 14 catégories listées dans le Tableau 4.1. Pour ces expérimentations, nous avons aussi utilisé des techniques d'augmentation afin d'obtenir 400 échantillons pour chaque catégorie, créant ainsi un ensemble d'entraînement de 5600 échantillons (14 catégories x 400 échantillons). Le graphique de la Figure 4.3 présente les détails de cette étape pour les ensembles multiétiquettes.

Tableau 4.1 Catégories utilisées pour les expérimentations multiétiquettes

No Catégorie	Étiquettes	Nb Échantillons
2	Parallèle	276
4	Arche	33
8	Hachure	361
10	Parallèle et hachure	14
16	Circulaire	124
18	Parallèle et circulaire	19
32	Granulaire	344
34	Parallèle et granulaire	426
36	Arche et granulaire	41
40	Hachure et granulaire	387
42	Parallèle, hachure et granulaire	110
44	Arche, hachure et granulaire	10
48	Circulaire et granulaire	140
64	Lisse	26

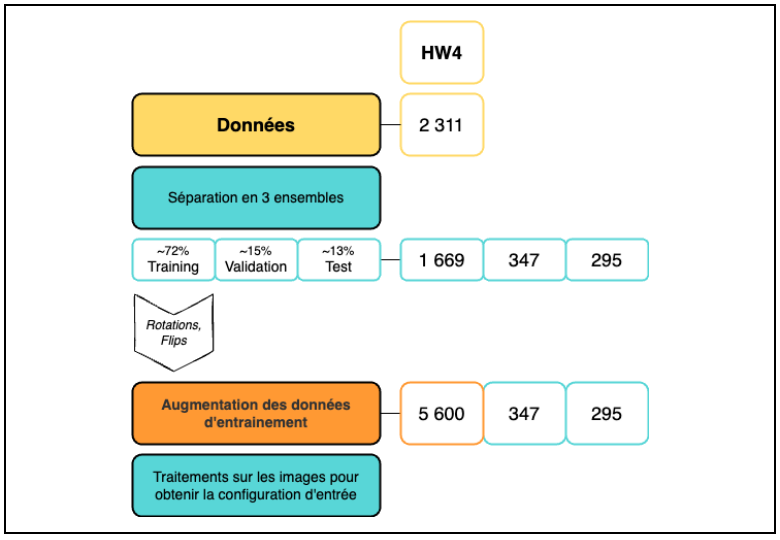


Figure 4.3 Division et augmentation des données pour les expérimentations multiétiquettes

Pour les classificateurs binaires, nous avons utilisé les données provenant du BrassTRAX HW4. Pour chacune des catégories, nous avons créé un ensemble d'échantillons incluant des marques de cette catégorie, avec une ou plusieurs étiquettes. Nous y avons ensuite ajouté le même nombre d'échantillons ne contenant pas de marques de cette catégorie. Puis, nous avons créé des étiquettes binaires (1 lorsque les marques sont présentes et 0 pour les autres). Pour terminer, nous avons utilisé des techniques d'augmentation avec un facteur différent pour chaque ensemble, selon la quantité d'échantillons de cette classe. Ainsi, les catégories avec moins d'échantillons, ont bénéficié d'un facteur d'augmentation plus élevé. Le diagramme de la Figure 4.4 illustre les étapes de préparation des données pour la classification binaire des marques parallèle et arche. Le Tableau 4.2 présente le détail de l'ajout d'échantillons et le Tableau 4.3 présente le détail de la division des données pour toutes les catégories.

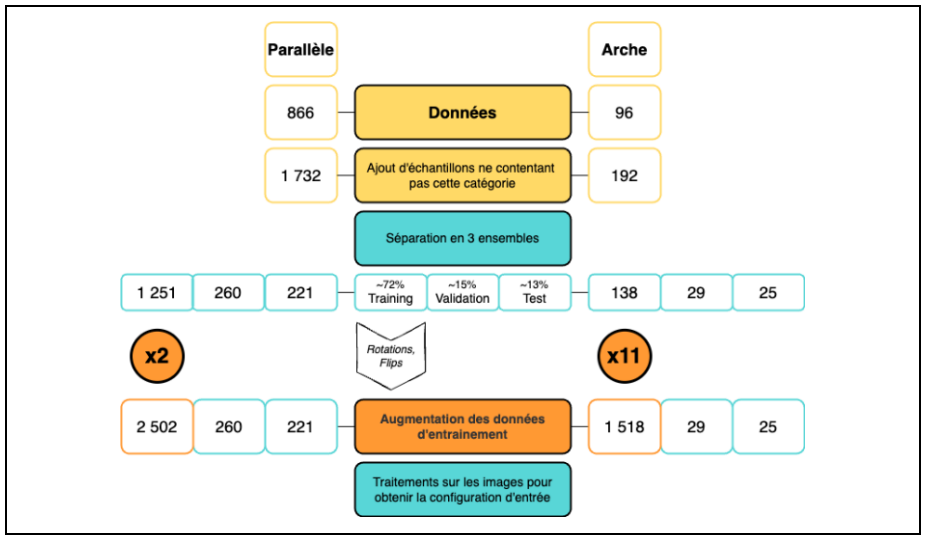


Figure 4.4 Division des données pour les classificateurs binaires

Tableau 4.2 Détails de l'ajout d'échantillons pour les classificateurs binaires

Catégorie	**Nb d'échantillons**	**Échantillons ajoutés**	**Total**
Parallèle	866	866	1732
Arche	96	96	192
Hachure	908	908	1816
Circulaire	315	315	630
Granulaire	1487	255*	1742
Lisse	36	36	72

** Ce chiffre n'est pas doublé comme pour les autres catégories, car il y avait moins d'échantillons non granulaires disponibles.*

Tableau 4.3 Détails de la division des données pour les classificateurs binaires

Catégorie	Entraînement	Validation	Test	Facteur augmentation	Entraînement augmenté
Parallèle	1251	260	221	2	2502
Arche	138	29	25	11	1518
Hachure	1311	273	232	2	2622
Circulaire	454	95	81	6	2724
Granulaire	1258	262	222	2	2516
Lisse	51	11	10	16	816

4.2.5 Implémentation

Pour l'implémentation, nous avons utilisé les modèles de CNN provenant de la librairie Keras: VGG (Keras, [s d]), EfficientNet (Keras, [s d]), ainsi que les modèles de ViT B16, L16, B32 et L32 provenant de la librairie vit-Keras (Morales, [s d]). Le tableau ci-dessous présente quelques spécifications pour les différentes architectures de ViT.

Tableau 4.4 Spécifications des modèles ViT

Modèle ViT	*B16*	*L16*	*B32*	*L32*
Taille du patch	16×16	16×16	32×32	32×32
Têtes d'attention	12	16	12	16
Profondeur de l'encodeur	12	24	12	24

Nous avons effectué plusieurs expérimentations préliminaires afin de choisir certains hyperparamètres de départ. Par exemple, pour les CNN, nous avons testé les architectures VGG16 et VGG19, ainsi que les architectures EfficientNet B3, B5 et B7. Nous avons ainsi sélectionné les architectures VGG16 et EfficientNet B3. Le détail et les résultats de ces expérimentations préliminaires ne sont pas présentés dans cette thèse.

Pour les modèles multiclasses, nous avons effectué des apprentissages par transfert en utilisant les poids provenant du modèle préentraîné sur la base de données ImageNet. Pour ce faire, nous avons utilisé un paramètre du modèle afin de geler toutes ses couches. Cette étape fixe les poids déjà appris, ce qui prévient leur modification lors de l'entraînement avec de nouvelles données. Nous avons ensuite remplacé la tête du modèle, c'est-à-dire la section composée de couches denses suivant la couche d'aplatissement (*flatten*) et se terminant par la couche de prédiction. La nouvelle tête est constituée de trois couches denses de dimension 128, 64 et 32, avec une fonction d'activation GELU. Des couches de dropout et de normalisation du lot (*batch normalisation*) s'insèrent entre ces couches denses. La couche de sortie, de dimension 6 pour les six classes recherchées, utilise une fonction d'activation softmax. Cette fonction devra être jumelée à une perte d'entropie croisée catégorielle (*categorical crossentropy*) lors de l'entraînement du modèle, afin de traiter le problème multiclasse. La Figure 4.5 montre un diagramme de cette tête apprenante, utilisée pour tous les modèles d'entraînement par transfert. Pour les modèles de CNN (VGG16 et EfficientNet B3), nous avons effectué quelques expérimentations additionnelles afin de tester l'entraînement à partir de zéro. Pour ce faire, les modèles ont été implémentés tels quels, non figés et sans remplacement de la tête. Nous avons utilisé des poids de départs initialisés

aléatoirement. Le Tableau 4.5 résume les différents modèles utilisés pour cette partie des expérimentations.

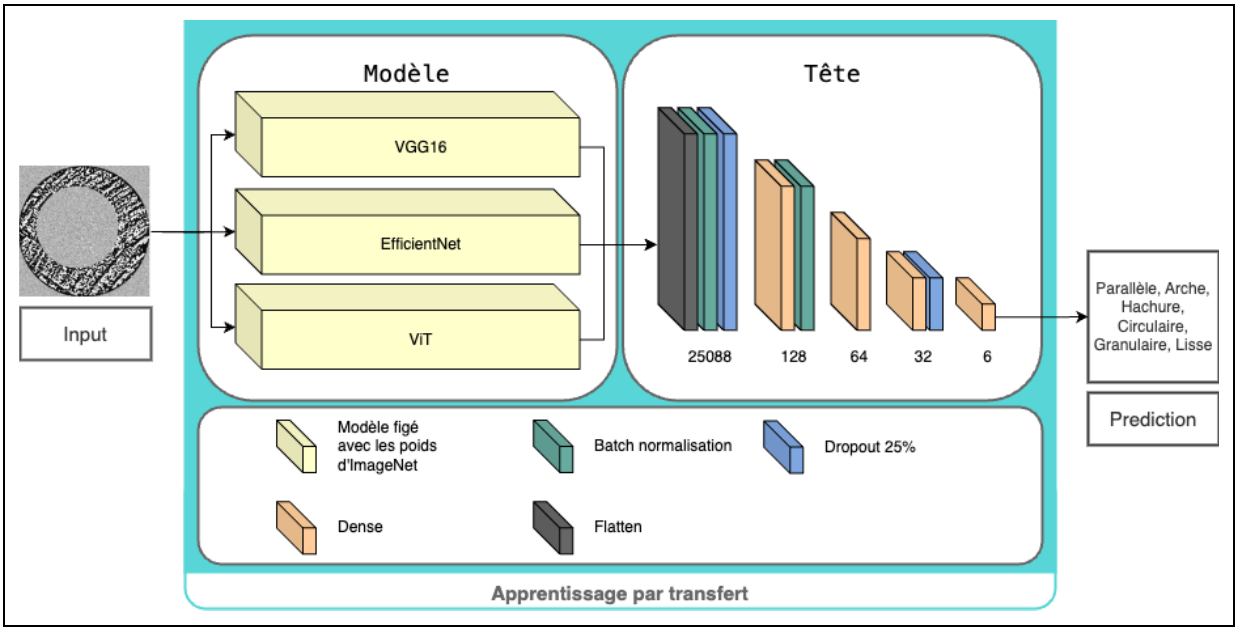


Figure 4.5 Diagramme de la tête apprenante pour l'entraînement par transfert

Tableau 4.5 Résumé des différents modèles multiclasses

	Données	***Apprentissage***	**VGG16**	**ENB3**	**ViT**
HW3	*Augmentées*	*Par transfert*	✓	✓	X
	Augmentées	*À partir de zéro*	✓	✓	X
HW4	*Augmentées*	*Par transfert*	✓	✓	✓
	Augmentées	*À partir de zéro*	✓	✓	X

Tableau 4.6 Résumé des différents modèles multiétiquettes

	Données	***Apprentissage***	**VGG16**	**ENB3**	**ViT B32**
HW4	*Augmentées*	*Par transfert*	✓	✓	✓
	Sans augmentation	*Par transfert*	✓	✓	✓

Pour les expérimentations multiétiquettes, nous avons seulement effectué des apprentissages par transfert en utilisant les modèles VGG16, EfficientNet B3 et ViT B32. Nous avons utilisé la tête présentée à la Figure 4.5, mais avec une fonction d'activation sigmoïde pour la couche de prédiction. Cette fonction d'activation, jumelée à une perte d'entropie croisée binaire plutôt que catégorielle, permet d'obtenir une valeur entre 0 et 1 pour chacune des catégories. Un seuil permet ensuite de déterminer individuellement si chaque catégorie est présente dans l'image. Nous avons débuté avec des entraînements à partir des données augmentées, et nous avons ensuite ajouté des entraînements à partir des données sans augmentation. Le Tableau 4.6 présente une liste des différents modèles entraînés pour cette partie des expérimentations.

Pour les expérimentations binaires, nous avons mis de côté l'architecture VGG16, car nous avons constaté le gain des autres architectures par rapport à cette référence dans les expérimentations précédentes. Puisque les modèles EfficientNet B3 se sont démarqués jusqu'à présent, nous les avons utilisés afin d'évaluer l'augmentation des données : en effectuant des apprentissages par transfert à partir des données augmentées, ainsi qu'à partir des données sans augmentation. Nous avons utilisé la tête présentée à la Figure 4.5, avec une fonction d'activation sigmoïde pour la couche de prédiction et une perte d'entropie croisée binaire lors de l'entraînement. Ensuite, nous avons ajouté les architectures EfficientNet B5 et B7, afin d'évaluer si leurs performances dépasseraient celles des modèles EfficientNet B3, ainsi que les architectures ViT B32 et L32. Nous avons repris les entraînements par transfert à partir des données non augmentées seulement. Le Tableau 4.7 présente une liste des différents modèles entraînés pour cette partie des expérimentations.

Tableau 4.7 Résumé des différents modèles binaires

	Données	***Apprentissage***	**ENB3**	**ENB5**	**ENB7**	**ViT B32**	**ViT L32**
HW4	*Augmentées*	*Par transfert*	✓	X	X	X	X
	Sans augmentation	*Par transfert*	✓	✓	✓	✓	✓

L'optimiseur Adam a été choisi pour entraîner tous les modèles. Les entraînements ont utilisé une taille du lot (batch size) de 32, avec une fonctionnalité d'arrêt anticipé gérée par le paramètre « callback ». Cet arrêt anticipé surveille l'exactitude de l'ensemble de validation avec un critère d'arrêt fixé à 25 épochs, le nombre d'épochs qui se sont déroulées sans amélioration de l'exactitude de validation. Pour l'évaluation des modèles et les matrices de confusion, nous avons utilisé les outils de la librairie métriques de *sklearn* (Scikit-learn, [s d]). Pour la génération des cartes thermographiques, nous avons utilisé la librairie *tf-keras-vis* (Kubota, [s d]). La génération des cartes d'attention est incluse dans la librairie *vit-Keras*.

4.3 Résultats

Cette section présente les résultats obtenus lors de l'entraînement des modèles et de leur évaluation avec les ensembles de test. Pour les entraînements, nous avons observé les courbes de performance et d'optimisation. Pour les évaluations, nous avons compilé l'exactitude et la perte, ainsi que les moyennes macros pour la précision, le rappel, la mesure-F1.

4.3.1 Classification multiclasse

La Figure 4.6 présente les courbes d'apprentissage des modèles CNN, entraînés avec les données augmentées provenant du BrassTRAX HW3. L'entraînement par transfert s'est déroulé sur approximativement 55 épochs, pour les deux modèles.

On peut observer que les courbes de performance des apprentissages par transfert pour VGG16 (Figure 4.6 c) et ENB3 (Figure 4.6 d) sont très proches l'une de l'autre. La courbe d'entraînement atteint rapidement une valeur dans le voisinage de 1 et demeure stable tout au long de l'entraînement. Cependant, la courbe de validation ne semble pas s'améliorer avec l'entraînement et demeure une ligne plus ou moins droite se situant autour de 60% pour VGG16 et 70% pour ENB3. Visuellement, nous constatons que ces courbes de performance ne convergent pas. En observant les courbes d'optimisation des deux mêmes modèles (Figure 4.6 g et h), nous remarquons que la courbe d'entraînement chute rapidement dans le

voisinage de zéro, alors que plus les épochs avancent, plus la courbe de validation augmente et l'écart entre les deux s'accroît. Ce qui nous indique que les modèles sont en état de surapprentissage. Ils tendent à apprendre des détails de l'ensemble d'entraînement qui ne sont pas pertinents à la classe représentée ni généralisable à l'ensemble de validation, tels que le bruit de l'image, ou des caractéristiques spéciales d'un échantillon particulier. Généralement, il s'agit d'une indication que les données d'entraînement sont insuffisantes, que les classes ne sont pas représentées significativement parmi les échantillons, ou que le modèle est trop complexe. Puisque les résultats semblent s'améliorer en augmentant la complexité du modèle, et que le nombre d'échantillons devrait être suffisant pour un apprentissage par transfert, nous considérons que les données d'entraînement ne représentent pas correctement les classes recherchées.

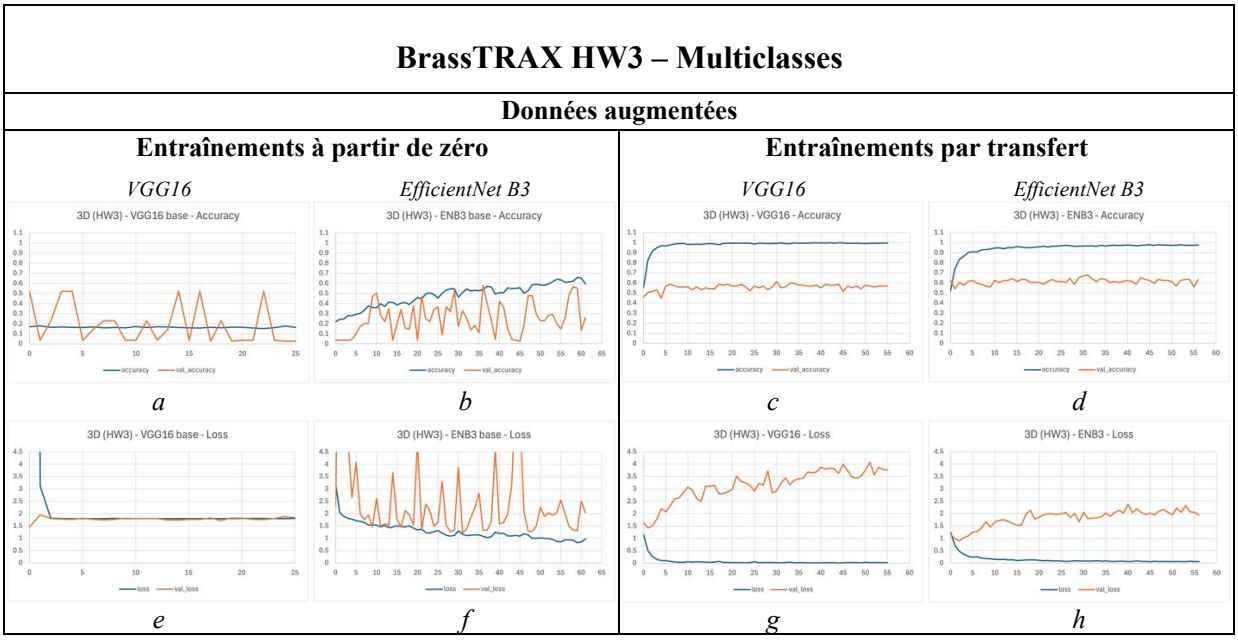


Figure 4.6 Courbes d'apprentissage des modèles CNN multiclasses, données augmentées du BrassTRAX HW3 : courbes de performance (haut) et courbes d'optimisation (bas)

En ce qui concerne les courbes d'apprentissage pour les modèles entraînés à partir de zéro, nous les avons jugées insatisfaisantes. Non seulement l'exactitude atteinte par l'ensemble d'entraînement sur les courbes de performance (Figure 4.6 a et b) est décevante (environ 20% pour VGG16 et moins de 60% pour ENB3), mais la courbe de l'ensemble de validation

démontre une grande instabilité, ce qui pourrait être une indication supplémentaire d'un sur apprentissage. Le phénomène se répète sur les courbes d'optimisation (Figure 4.6 e et f), avec une perte élevée qui ne descend pas sous 2 pour VGG16 et sous 1 pour ENB3. Ces courbes d'apprentissage nous indiquent que les modèles ne parviennent pas à apprendre correctement à partir des données disponibles.

À la suite de ces expérimentations, nous avons effectué une deuxième série d'entraînements des modèles de CNN en utilisant les données augmentées provenant du BrassTRAX HW4. La Figure 4.7 présente les courbes d'apprentissage associées à ces essais. L'entraînement par transfert s'est déroulé sur 43 époch pour le modèle VGG16 et 58 époch pour le modèle ENB3.

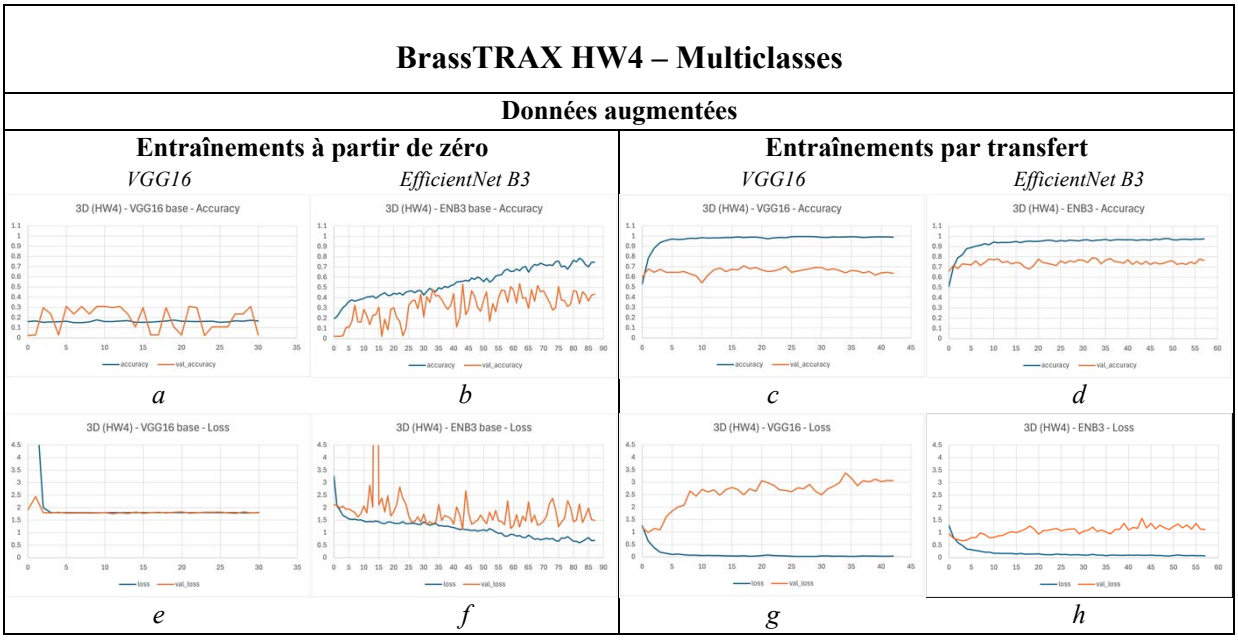


Figure 4.7 Courbes d'apprentissage des modèles CNN multiclasses, données augmentées du BrassTRAX HW4: courbes de performance (haut) et courbes d'optimisation (bas)

Pour les apprentissages par transfert, nous remarquons que les courbes de validation obtenues sur le modèle ENB3 (Figure 4.7 d et h) semblent se rapprocher davantage des courbes d'entraînement que celles obtenues sur le modèle VGG16 (Figure 4.7 c et g), en atteignant presque 80,00% d'exactitude. Cependant, la courbe de perte de validation du modèle ENB3

tend vers 1,00 et semble vouloir continuer à augmenter avec les époch. Pour le modèle VGG16, on observe une exactitude de validation s'approchant de 70,00% et une perte de validation dépassant 3,00. Nous observons tout de même des performances supérieures avec les modèles qui sont entraînés à partir des données provenant du HW4. Nous expliquons ce phénomène par le fait que, même si le nombre total d'échantillons provenant du HW4 est moins élevé que le nombre d'échantillons provenant du HW3 (3609 contre 9822), le nombre d'échantillons avec une seule étiquette provenant du HW4 est plus élevé (1164 contre 724). Il est aussi possible que les étiquettes de la vérité de terrain des données provenant du HW4 soient plus représentatives des classes présentes sur les images.

Pour les entraînements à partir de zéro (Figure 4.7 a, b, e et f), nous observons la même tendance décrite précédemment pour les modèles HW3, avec des valeurs légèrement supérieures pour les courbes de performances et légèrement inférieures pour les courbes d'optimisation. Mais malgré ces améliorations, notre intérêt pour ces modèles demeure limité, dû aux faibles valeurs d'exactitude, aux valeurs élevées de perte, et à l'instabilité des courbes de validation. Ces observations nous poussent à mettre de côté les modèles entraînés à partir de zéro et à nous concentrer sur l'entraînement par transfert pour la suite des expérimentations. Nous croyons que le nombre d'échantillons à notre disposition n'est pas suffisamment élevé pour permettre d'entraîner efficacement un modèle à partir de zéro pour ces tâches de classification.

Pour la dernière série d'expérimentations multiclasses, nous avons entraîné quatre modèles de ViT avec les données augmentées provenant du BrassTRAX HW4. La Figure 4.8 présente les courbes d'apprentissage associées à ces essais. Nous observons que le modèle ViT B32 (Figure 4.8 c et g) semble le plus intéressant, car les courbes d'entraînement et de validation sont celles qui se rapprochent le plus l'une de l'autre, indiquant une meilleure capacité de généralisation. Malgré cela, l'exactitude d'entraînement stagne à 80,00%, indiquant que le modèle n'apprend pas assez avec les données d'entraînement. L'entraînement s'est déroulé sur 44 époch et la perte de validation ne descends (presque) pas en dessous de 1,00. Les courbes d'entraînement et de validation divergent l'une de l'autre, indiquant que des époch

supplémentaires ne causeraient qu'un surentraînement. Pour les courbes d'apprentissage des trois autres modèles de ViT (B16, L16 et L32), nous observons des comportements similaires, mais atteignant des valeurs légèrement inférieures (c.-à-d. exactitude plus faible, perte plus élevée). Bref, en se basant sur les courbes d'apprentissage, les modèles de ViT ne parviennent pas à surpasser le modèle ENB3 présenté plus haut.

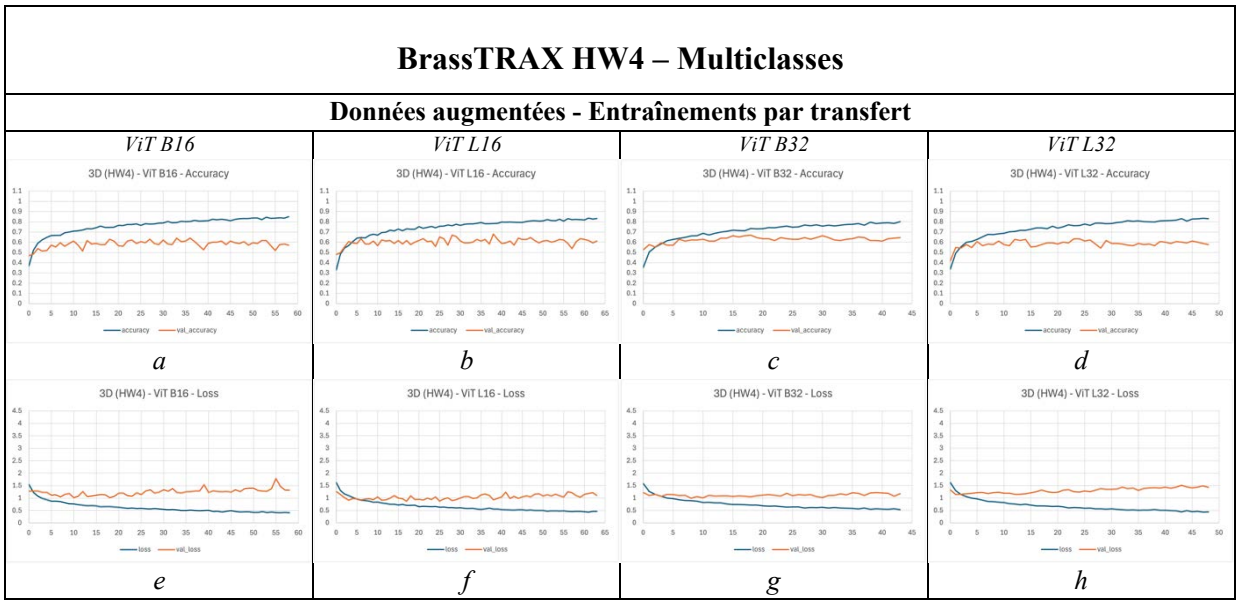


Figure 4.8 Courbes d'apprentissage des modèles ViT multiclasses, données augmentées du BrassTRAX HW4 : courbes de performance (haut) et courbes d'optimisation (bas)

Nous avons ensuite utilisé les ensembles de test pour évaluer les modèles. Les tableaux ci-dessous présentent les métriques pour les modèles entraînés à partir des données HW3 (Tableau 4.8) et HW4 (Tableau 4.9). Les graphiques de la Figure 4.9 présentent une comparaison visuelle de ces mêmes informations. Les modèles entraînés à partir de zéro sont identifiés par le terme -base. Nous observons que toutes les métriques sont plus élevées pour le modèle EfficientNet B3, entraîné par transfert. Pour la perte, ce sont les ViT qui obtiennent les valeurs les plus basses, mais la perte pour le modèle ENB3 demeure acceptable.

Tableau 4.8 Métriques d'évaluation des modèles entraînés avec les données HW3

HW3	Exactitude	Précision	Rappel	Mesure-F1	Perte
VGG16-base	52,69%	8,78%	16,67%	11,50%	1,46
ENB3-base	54,84%	42,42%	23,96%	23,68%	1,23
VGG16	54,84%	45,60%	53,20%	47,30%	3,20
ENB3	69,89%	71,72%	67,40%	64,49%	1,67

Tableau 4.9 Métriques d'évaluation des modèles entraînés avec les données HW4

HW4	Exactitude	Précision	Rappel	Mesure-F1	Perte
VGG16-base	30,87%	5,15%	16,67%	7,86%	1,78
ENB3-base	55,03%	42,69%	37,08%	37,84%	1,27
VGG16	67,79%	51,52%	58,95%	53,28%	1,76
ENB3	77,85%	74,94%	69,73%	71,53%	1,20
ViT B16	61,74%	48,63%	57,08%	50,42%	1,22
ViT L16	65,10%	66,04%	68,37%	63,49%	0,98
ViT B32	57,05%	47,72%	57,34%	49,10%	1,07
ViT L32	69,80%	72,62%	62,90%	65,80%	0,75

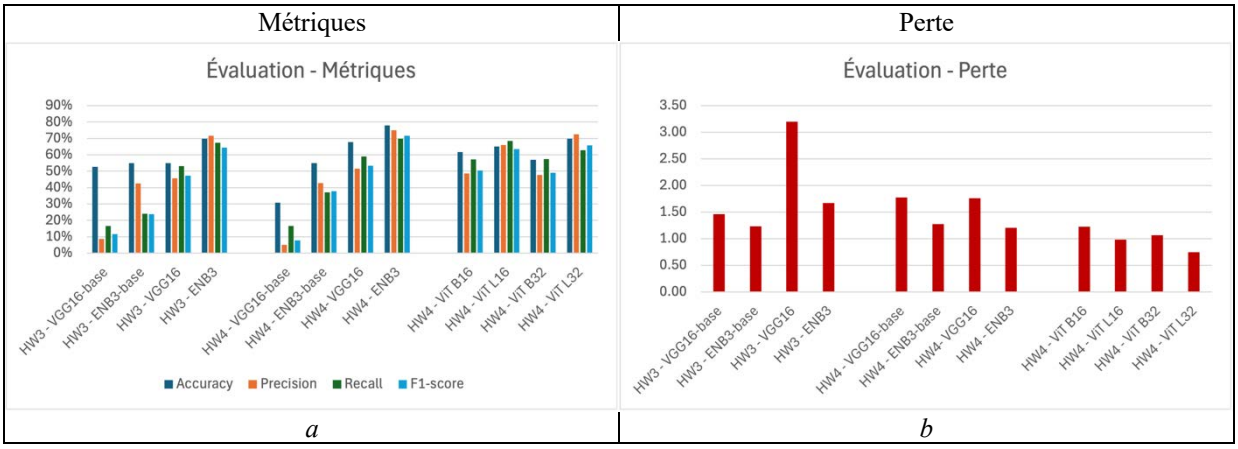


Figure 4.9 Graphiques comparatifs des métriques d'évaluation (a) et de la perte (b) pour tous les modèles (HW3 et HW4)

Nous avons ensuite compilé les matrices de confusion pour chacun des modèles. Ces matrices, présentées à la Figure 4.10, nous permettent d'observer le comportement des modèles lors des classifications. La dimension des graphiques est réduite afin qu'ils puissent tous être affichés sur une même page, ce qui rend les chiffres difficiles à lire. Mais pour le besoin d'illustration, les couleurs sont suffisantes : l'échelle de couleur passe du foncé (aucun échantillon) au très clair (le plus grand nombre d'échantillons, c'est-à-dire 40 dans ce cas). L'axe des x correspond à la prédiction du modèle, alors que l'axe des y correspond à la vérité de terrain de l'étiquette.

Figure 4.10 Matrices de confusion multiclasses pour tous les modèles

Pour les modèles VGG16 entraînés à partir de zéro (Figure 4.10 a et e), tous les échantillons ont reçu la même prédiction, qui correspond à la classe majoritaire de l'ensemble de test (granulaire pour HW3 et hachure pour HW4). Pour le modèle ENB3 entraîné à partir de zéro sur HW3 (Figure 4.10 b), il émet cinq prédictions originales, qui ne correspondent pas à la classe majoritaire. Quant au modèle ENB3 entraîné à partir de zéro sur HW4 (Figure 4.10 f), on peut observer que plusieurs prédictions s'écartent de la classe majoritaire, et que plusieurs prédictions sont correctes, même si dans l'ensemble les résultats sont décevants. Ces prédictions médiocres sont aussi reflétées par les métriques, présentées précédemment. Malgré une exactitude de 52,69% et 30,87% pour les deux modèles VGG16 entraînés à partir de zéro, les autres métriques demeurent basses: par exemple 11,50% et 7,86% pour la mesure-F1. Ces indications appuient ce que nous avons observé dans les courbes d'apprentissage pour ces modèles et confirment notre décision d'abandonner les entraînements à partir de zéro.

Les autres modèles présentent des matrices de confusion plus intéressantes. D'abord, on peut y observer des prédictions plus balancées. On peut aussi constater que les modèles entraînés à partir des données HW4 offrent un nombre plus élevé de prédictions correctes : en observant la diagonale représentant les vrais positifs (TP) on remarque que les carrés sont de couleur plus pâle, indiquant un nombre plus élevé. On peut observer que le modèle ENB3, entraîné par transfert à partir des données HW4 (Figure 4.10 h), présente la diagonale la plus claire, avec le moins de zones claires à l'extérieur de la diagonale (indiquant les erreurs de prédiction). Ces prédictions bonifiées sont aussi reflétées par les métriques présentées plus haut pour ce modèle.

Finalement, nous avons tenté de mieux comprendre les résultats de classification des CNN en observant les cartes thermographiques obtenues à la sortie des modèles par la technique GradCam ++. Pour les modèles de ViT, nous avons observé les cartes d'attention. Les résultats de visualisation pour les modèles multiclasses sont présentés en annexe (ANNEXE I, Figure-A I-1). Nous remarquons que dans plusieurs cas, le modèle ne semble pas porter son attention au bon endroit. Par exemple, on peut observer que les zones importantes pour le

réseau se situent un peu n'importe où sur l'image. Le modèle a appris des caractéristiques qui sont propres aux images d'entraînement et qui ne correspondent pas au résultat recherché. On remarque aussi que les cartes thermographiques pour ENB3 et les cartes d'attention pour ViT B32 semblent parvenir à se concentrer davantage sur la région qui nous intéresse, c'est-à-dire la marque de la face de la culasse.

De façon générale, l'absence de convergence dans les courbes d'apprentissage, notamment avec des courbes de performance de validation, parallèles aux courbes d'entraînement, mais décalées de 10% ou plus; ainsi que des courbes d'optimisation, où la courbe de validation s'éloigne de la courbe d'entraînement au fil des époch, nous indiquent que nos données ne sont pas parfaitement appropriées à la tâche demandée. À cette étape des expérimentations, nous nous sommes questionnés sur la justesse des étiquettes de la vérité de terrain. Nous avons commencé par une évaluation visuelle des échantillons provenant du HW3, et nous avons constaté plusieurs discordances et erreurs parmi les étiquettes. Par exemple, deux images d'échantillons à l'apparence hautement similaire qui sont étiquetés différemment, ou des images exhibant clairement un type de marque (circulaire par exemple), mais étant étiquetés avec d'autres catégories. Nous avons aussi remarqué qu'un grand nombre d'échantillons étiquetés comme étant inconnus auraient pu être étiquetés avec au moins une des six catégories de marque. À la suite de ces observations, nous avons demandé l'opinion d'un second expert. Cet exercice prend la forme d'un accord interannotateur préliminaire, et les observations sont présentées à la fin des résultats (section 4.3.4)

4.3.2 Classification multiétiquette

La Figure 4.11 présente les courbes d'apprentissage obtenues lors des expérimentations à partir des données augmentées, avec les trois modèles sélectionnés pour cette partie des expérimentations.

Nous remarquons que le modèle ENB3 (Figure 4.11 b et e) semble mal adapté pour les données multiétiquettes. Tout d'abord, la courbe de performance pour l'entraînement, qui

n'atteint pas 25%, est au-dessous de la courbe de validation, indiquant un sous-entraînement : le modèle ne parvient pas à apprendre à partir des données qu'on lui a fournies. Pour le modèle VGG16 (Figure 4.11 a et d), les formes des courbes sont semblables à ce que nous avons observé précédemment, avec des valeurs d'exactitude moins élevées, indiquant une performance inférieure. Cependant, la courbe d'optimisation de l'ensemble de validation atteint une perte plus basse, mais s'éloigne toujours de la courbe d'entraînement, indiquant un surapprentissage du modèle. Quant au modèle ViT B32 (Figure 4.11 c et f), il présente les courbes les plus intéressantes pour cette partie des expérimentations. Les courbes d'optimisation atteignent une perte de 0,5, ce qui est préférable à ce que nous avons obtenu lors des expérimentations multiclasses. Toutefois, le niveau d'exactitude des courbes de performance demeure bas. De plus, aucune des courbes ne converge et la courbe d'optimisation de validation s'éloigne de la courbe d'entraînement au fil des épochs, indiquant encore une fois un sur apprentissage.

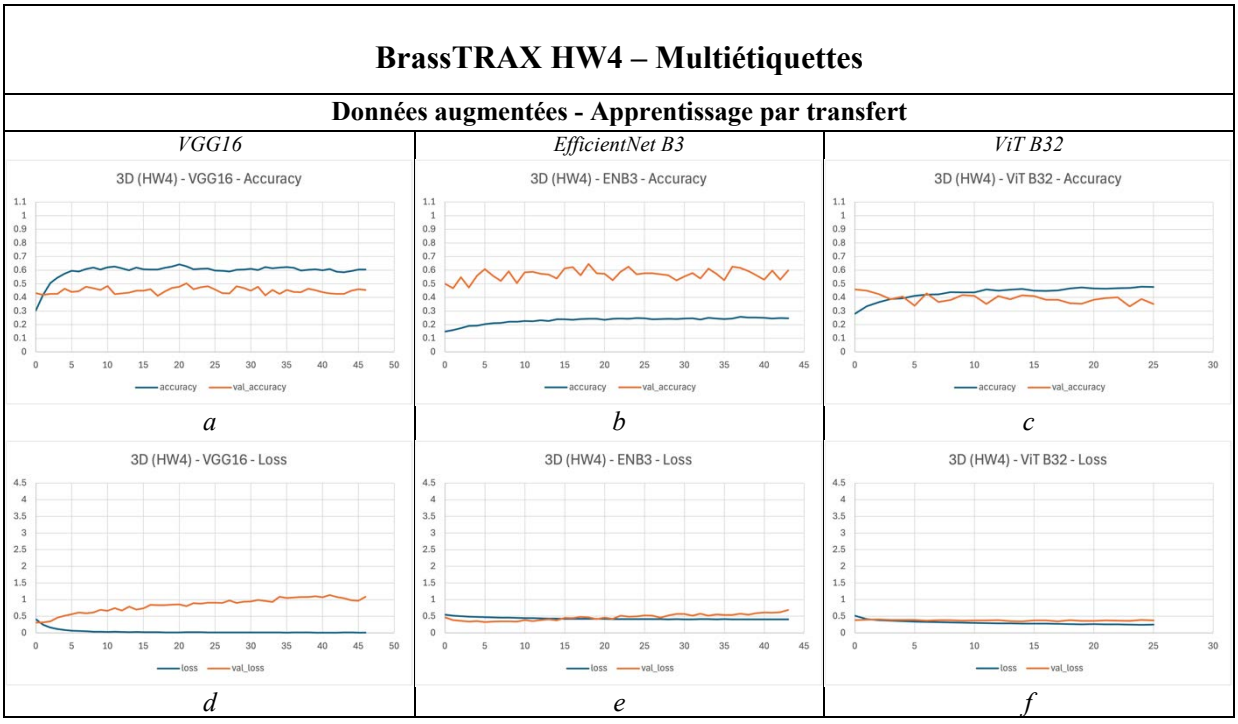


Figure 4.11 Courbes d'apprentissage des modèles multiétiquettes, données augmentées du BrassTRAX HW4: courbes de performance (haut) et courbes d'optimisation (bas)

À la suite de ces résultats, nous avons repris les mêmes expérimentations avec les données sans augmentation. L'augmentation que nous avons appliquée, vise à équilibrer l'ensemble de données en sur augmentant les catégories moins présentes. Dans le cas des données multiétiquettes, c'est au niveau des combinaisons de catégories que cet ajustement s'effectue et non au niveau des catégories de base. De cette façon, nous avons augmenté chaque combinaison de catégories à 400 échantillons. Ainsi, la combinaison 34, qui a 308 échantillons dans l'ensemble d'entraînement, a été augmentée aléatoirement 92 fois, alors que la combinaison 44, qui a huit échantillons dans l'ensemble d'entraînement, a été augmentée aléatoirement 392 fois, ce qui nous semble excessif et pourrait expliquer le comportement étrange de la courbe ENB3. De plus, le nombre d'échantillons dans les catégories de base (les six types de marques) n'a pas été ajusté, et demeure hautement déséquilibré : par exemple, sept combinaisons contiennent la marque granulaire alors qu'une seule contient la marque lisse (voir le Tableau 4.1 pour les combinaisons de catégories). La Figure 4.12 présente les courbes d'apprentissage obtenues lors des expérimentations avec les données sans augmentation.

Ces nouvelles courbes sont beaucoup plus intéressantes. Particulièrement celles du modèle EfficientNet B3 (Figure 4.12 b et e). Même si les courbes ne convergent toujours pas, on peut observer que les courbes de validation sont proches des courbes d'entraînement, indiquant une meilleure capacité à généraliser, et que les courbes d'optimisation atteignent des niveaux plus bas pour la perte. Nous croyons que ces métriques pourraient être améliorées en augmentant l'ensemble sans tenir compte du déséquilibre (en augmentant tous les échantillons de la même quantité), ou en utilisant une fonction de perte qui tient compte du déséquilibre, telle que la perte focale.

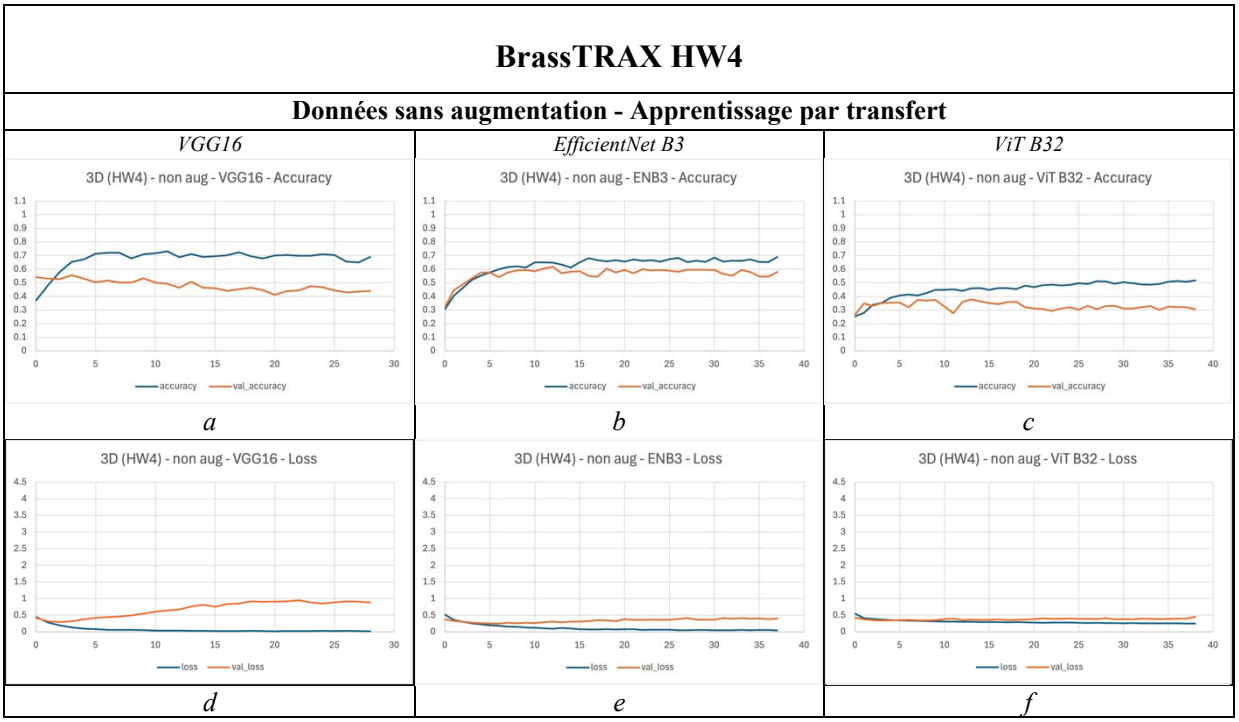


Figure 4.12 Courbes d'apprentissage des modèles multiétiquettes, données sans augmentation du BrassTRAX HW4: courbes de performance (haut) et courbes d'optimisation (bas)

Tableau 4.10 Métriques d'évaluation des modèles multiétiquettes

HW4	Exactitude	Précision	Rappel	Mesure-F1	Perte
VGG16 (aug)	44,07%	51,37%	57,53%	53,18%	0,92
ENB3 (aug)	53,22%	54,57%	54,29%	53,09%	0,52
ViTB32 (aug)	41,36%	42,32%	46,07%	43,14%	0,41
VGG16	52,54%	61,70%	48,08%	52,60%	0,36
ENB3	52,88%	55,78%	54,24%	54,47%	0,36
ViT B32	31,53%	59,92%	44,96%	44,16%	0,38

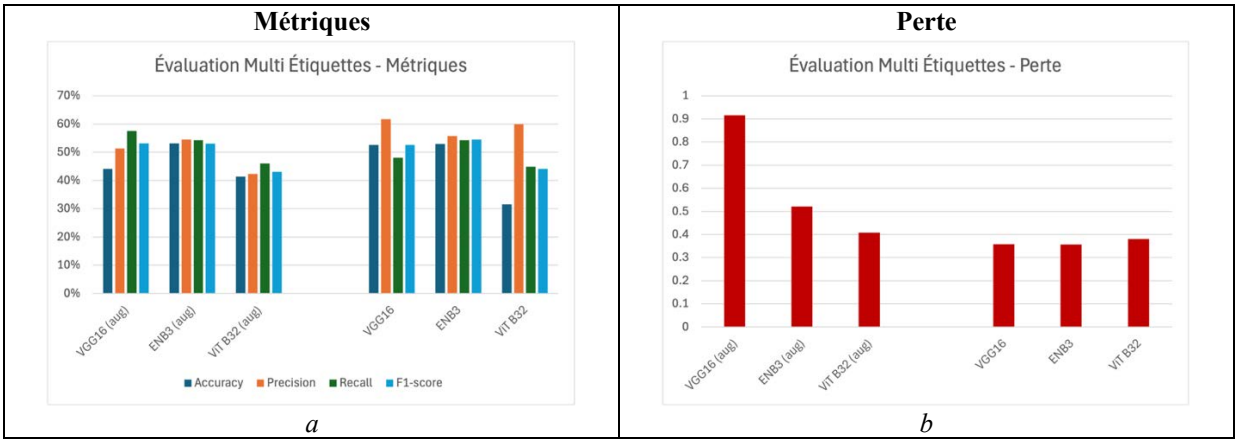


Figure 4.13 Graphiques comparatifs des métriques d'évaluation (a) et de la perte (b) pour tous les modèles multiétiquettes

Le Tableau 4.10 présente les métriques obtenues pour tous les modèles, alors que les graphiques de la Figure 4.13 présentent une comparaison visuelle de ces mêmes informations. Nous observons que les valeurs plus élevées sont partagées parmi les modèles. Le modèle ENB3 (aug) obtient la valeur la plus élevée pour l'exactitude, VGG16 pour la précision, VGG16 (aug) pour le rappel et ENB3 pour la mesure-F1. Les valeurs de perte les plus basses sont pour les modèles entraînés à partir des données non augmentées. En se basant sur les courbes d'apprentissage et les métriques, le modèle ENB3 (sans augmentation des données) serait le modèle le plus performant pour ces expérimentations. Nous avons ensuite compilé les matrices de confusion binaire pour chacun des modèles. Ces matrices binaires, correspondant aux six catégories, sont présentées à la Figure 4.14 pour le modèle EfficientNet B3. Les matrices binaires pour les autres modèles sont présentées en annexe (ANNEXE II, Figure-A II-1).

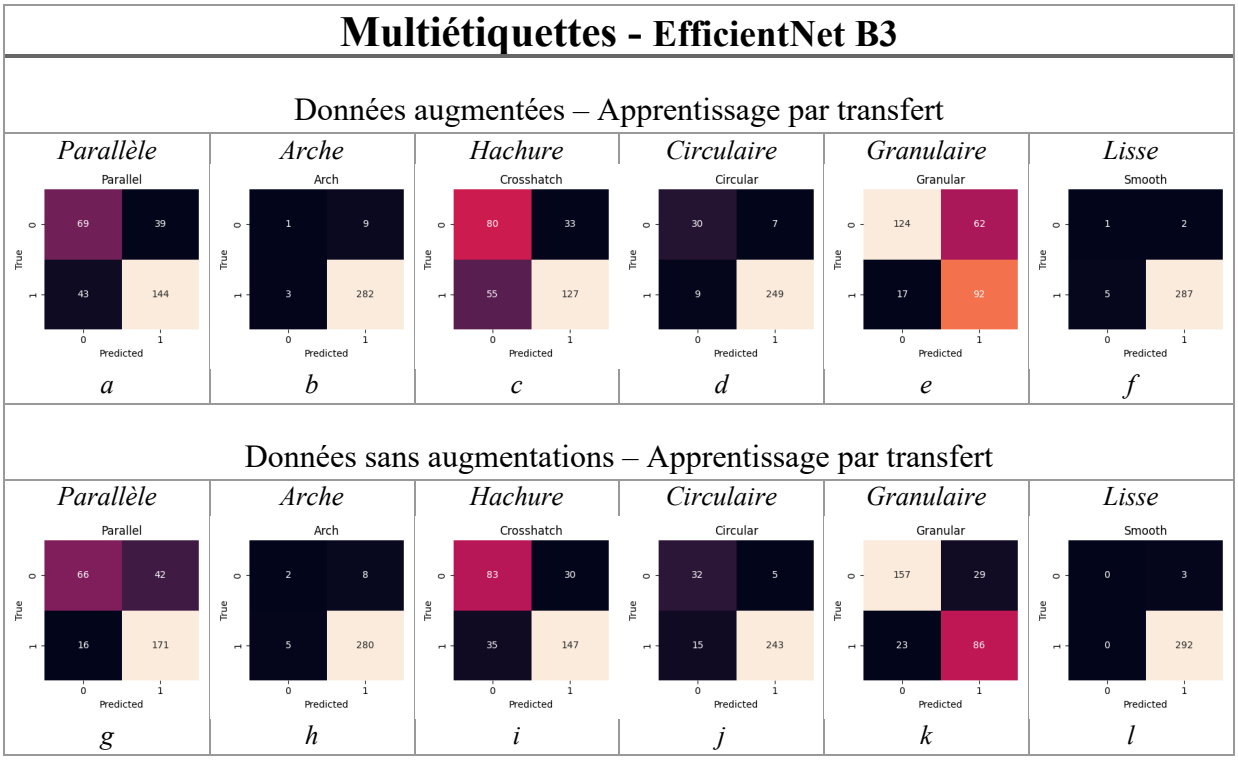


Figure 4.14 Matrices de confusion binaires pour les modèles EfficientNet B3 multiétiquettes : entraînés avec données augmentées (a-f) et sans augmentations (g-l)

L'interprétation des matrices de confusion est plus difficile pour les données multiétiquettes, car la performance d'un modèle peut varier selon la catégorie. On remarque que tous les modèles semblent avoir de la difficulté à prédire correctement les marques parallèle, hachure et granulaire. Nous arrivons à cette conclusion, car le nombre d'erreurs de prédiction (les faux positifs FP et les faux négatifs FN) est plus élevé pour ces catégories. Par exemple, la matrice de confusion pour la marque parallèle (Figure 4.14 a) indique 39 FP et 43 FN, pour un total de 82 erreurs sur 295 échantillons. Certaines matrices affichent un plus grand nombre d'erreurs FP (Figure 4.14 e et g), c'est-à-dire que le modèle identifie un type de marque qui n'est pas présent. Alors que d'autres obtiennent un plus grand nombre d'erreurs FN (Figure 4.14 c et j), c'est-à-dire que le modèle échoue à identifier un type de marque qui est là. Mais ces erreurs diffèrent selon que le modèle a été entraîné à partir des données augmentées ou sans augmentations, alors nous croyons que ces erreurs pourraient être dues à l'incohérence de l'étiquetage mentionné précédemment. De plus, il est difficile d'analyser le comportement pour les marques arche et lisse, car ce type de marque est sous-représenté

parmi les échantillons. Le faible taux d'erreurs est probablement dû à la quantité limitée des échantillons.

L'observation des cartes thermographiques et des cartes d'attention des modèles multiétiquettes, présentées en annexe (ANNEXE II, Figure-A II-2) confirme que certains modèles ont tendance à porter leur attention au mauvais endroit. On remarque peu de différences entre les cartes obtenues à partir des modèles entraînés avec les données augmentées et celles obtenues à partir des modèles entraînés avec les données sans augmentation, ce qui nous indique que les augmentations ont un effet sur la performance des modèles, mais pas sur l'identification de la région d'intérêt.

4.3.3 Classification binaire

Cette section présente les courbes d'apprentissage obtenues lors des expérimentations avec le modèle ENB3 pour les entraînements à partir des données augmentées (Figure 4.15) et à partir des données sans augmentation (Figure 4.16).

En observant ces courbes, nous remarquons qu'elles sont plus intéressantes que celles des expérimentations précédentes : les courbes de performances atteignent des valeurs plus élevées et les courbes d'optimisations, des valeurs plus basses. Toutefois, nous observons toujours une absence de convergence, et une courbe d'optimisation de validation qui s'éloigne de celle d'entraînement au fil des époch sur presque tous les modèles, indiquant encore une fois un sur entraînement. L'augmentation des données ne semble pas améliorer les courbes. Au contraire, elles semblent empirer pour les marques arche et lisse, les deux catégories possédant le moins d'échantillons originaux. Pour ces catégories, nous observons des courbes de validation qui s'éloignent davantage des courbes d'entraînement avec l'augmentation, autant pour les graphiques de performance (Figure 4.15 et Figure 4.16, b et f), que pour ceux d'optimisation (Figure 4.15 et Figure 4.16, h et l).

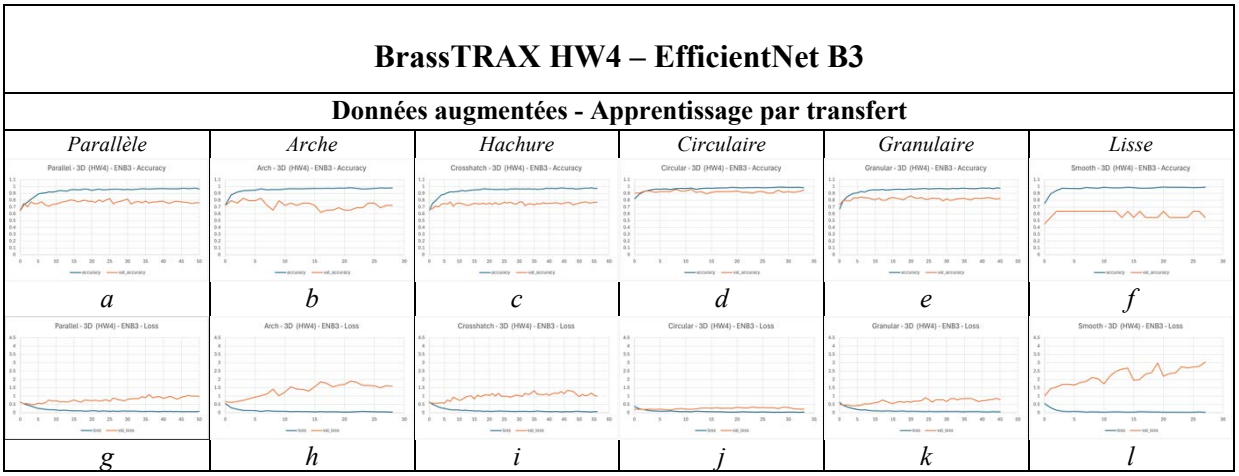


Figure 4.15 Courbes d'apprentissage du modèle binaire ENB3, données augmentées du BrassTRAX HW4 courbes de performance (haut) et courbes d'optimisation (bas)

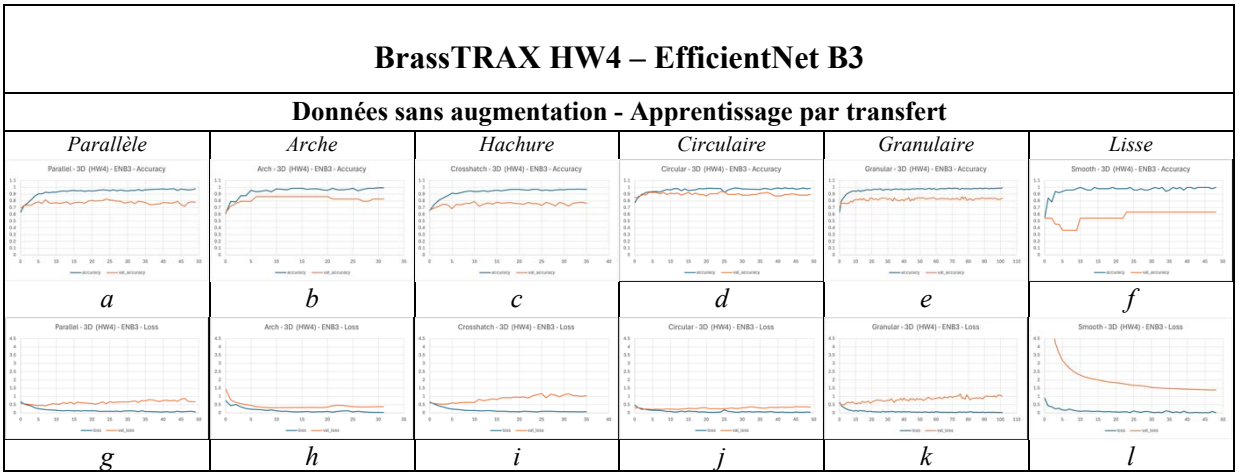


Figure 4.16 Courbes d'apprentissage du modèle binaire ENB3, données sans augmentation du BrassTRAX HW4: courbes de performance (haut) et courbes d'optimisation (bas)

En inspectant les courbes des modèles entraînés à partir des données sans augmentation, nous n'avons pas observé de différences significatives entre les trois modèles EfficientNet (B3, B5 et B7). Les modèles ViT (B32 et L32) semblent légèrement inférieurs, tout comme dans les expérimentations présentées précédemment. Les courbes d'entraînement de ces modèles sont disponibles en annexe (ANNEXE III, Figure-A III-1 à 4).

Tableau 4.11 Métriques d'évaluation des modèles binaires EfficientNet B3, pour les données augmentées

HW4 – ENB3							
Catégorie	**EX**	**PR**	**Rappel**	**F1**	**MCC**	**Auroc**	**Perte**
Parallèle	77,83%	77,89%	77,82%	77,81%	0,56	86,58%	0,80
Arche	80,00%	80,19%	79,81%	79,87%	0,60	89,74%	0,48
Hachure	75,43%	75,66%	75,43%	75,38%	0,51	83,14%	1,21
Circulaire	92,59%	92,67%	92,62%	92,59%	0,85	98,60%	0,19
Granulaire	81,53%	81,53%	81,53%	81,53%	0,63	88,64%	0,74
Lisse	60,00%	60,00%	60,00%	60,00%	0,20	64,00%	1,15

Tableau 4.12 Métriques d'évaluation des modèles binaires EfficientNet B3, pour les données sans augmentation

HW4 – ENB3							
Catégorie	**EX**	**PR**	**Rappel**	**F1**	**MCC**	**Auroc**	**Perte**
Parallèle	81,00%	81,51%	80,97%	80,91%	0,62	89,70%	0,67
Arche	88,00%	88,14%	88,14%	88,00%	0,76	94,23%	0,33
Hachure	78,88%	78,98%	78,88%	78,86%	0,58	83,76%	0,69
Circulaire	92,59%	92,59%	92,59%	92,59%	0,85	98,41%	0,22
Granulaire	79,73%	80,28%	79,73%	79,64%	0,60	88,10%	1,02
Lisse	70,00%	70,83%	70,00%	69,70%	0,41	64,00%	2,01

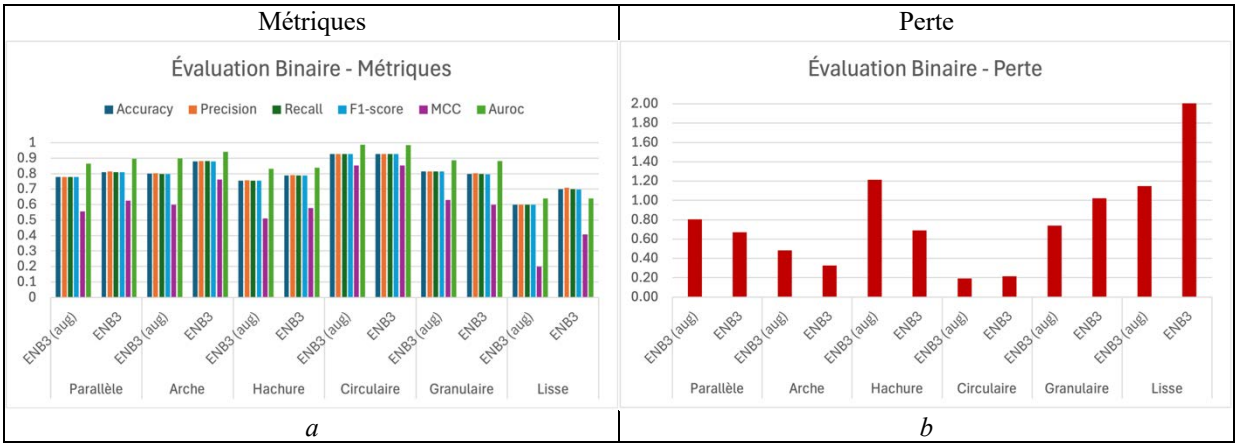


Figure 4.17 Graphiques comparatifs des métriques d'évaluation (a) et de la perte (c) pour les modèles binaires ENB3, avec et sans augmentation

Le Tableau 4.11 présente les métriques obtenues par les modèles ENB3 pour les données augmentées, et le Tableau 4.12 présente ces métriques pour les données sans augmentation. Deux métriques s'ajoutent à l'exactitude (EX), la précision (PR), le rappel et la mesure-F1: le coefficient de corrélation de Matthews (MCC) et l'aire sous la courbe RO (Auroc). Les graphiques de la Figure 4.17 présentent une comparaison visuelle des mêmes informations. Pour faciliter la comparaison visuelle, nous avons inclus la métrique MCC sur le graphique avec les autres métriques, même si son étendue est de [-1, 1]. Pour ce graphique, nous avons utilisé une échelle de [0, 1], car toutes les valeurs de MCC sont supérieures à zéro. Nous remarquons que les résultats d'évaluation sont variables. Pour les modèles des marques parallèle, arche, hachure et lisse, les résultats sont nettement supérieurs lorsque les modèles sont entraînés à partir des données sans augmentation. À l'inverse, pour les modèles de la marque granulaire, les résultats sont légèrement supérieurs lorsque les modèles sont entraînés à partir des données augmentées. Pour les modèles de la marque circulaire, nous n'observons presque pas de changement dans les résultats. Nous n'établissons pas de lien entre ces différences et le nombre de données originales disponibles pour les différentes catégories, et nous ne trouvons pas d'explication de ces résultats dans les courbes d'entraînement présentées précédemment. Nous supposons que la catégorie granulaire pourrait posséder un plus grand nombre d'échantillons avec des erreurs d'étiquetage, ou que l'augmentation aléatoire a pu sélectionner une majorité d'échantillons sans erreurs.

La Figure 4.18 présente les graphiques comparatifs de l'évaluation de tous les modèles entraînés à partir des données sans augmentation. Les tableaux présentant le détail de ces évaluations supplémentaires sont disponibles en annexe (ANNEXE III, Tableau-A III-1 à 6). Pour toutes les catégories, nous remarquons que les meilleurs résultats sont détenus par le modèle ENB3; sauf pour la catégorie circulaire, où il est surpassé par le modèle ENB5. Cependant, la perte varie beaucoup selon les catégories, avec des valeurs élevées pour la catégorie lisse, qui n'a presque pas d'échantillons dans l'ensemble. Nous remarquons que toutes les métriques sont assez stables entre elles, c'est-à-dire que lorsqu'une métrique est inférieure dans un modèle, toutes les autres métriques semblent baisser proportionnellement. Sauf pour la métrique MCC, qui baisse davantage lorsque la complexité du modèle augmente. De façon générale, les modèles de la catégorie circulaire obtiennent des métriques plus élevées et des pertes plus basses que les modèles des autres catégories.

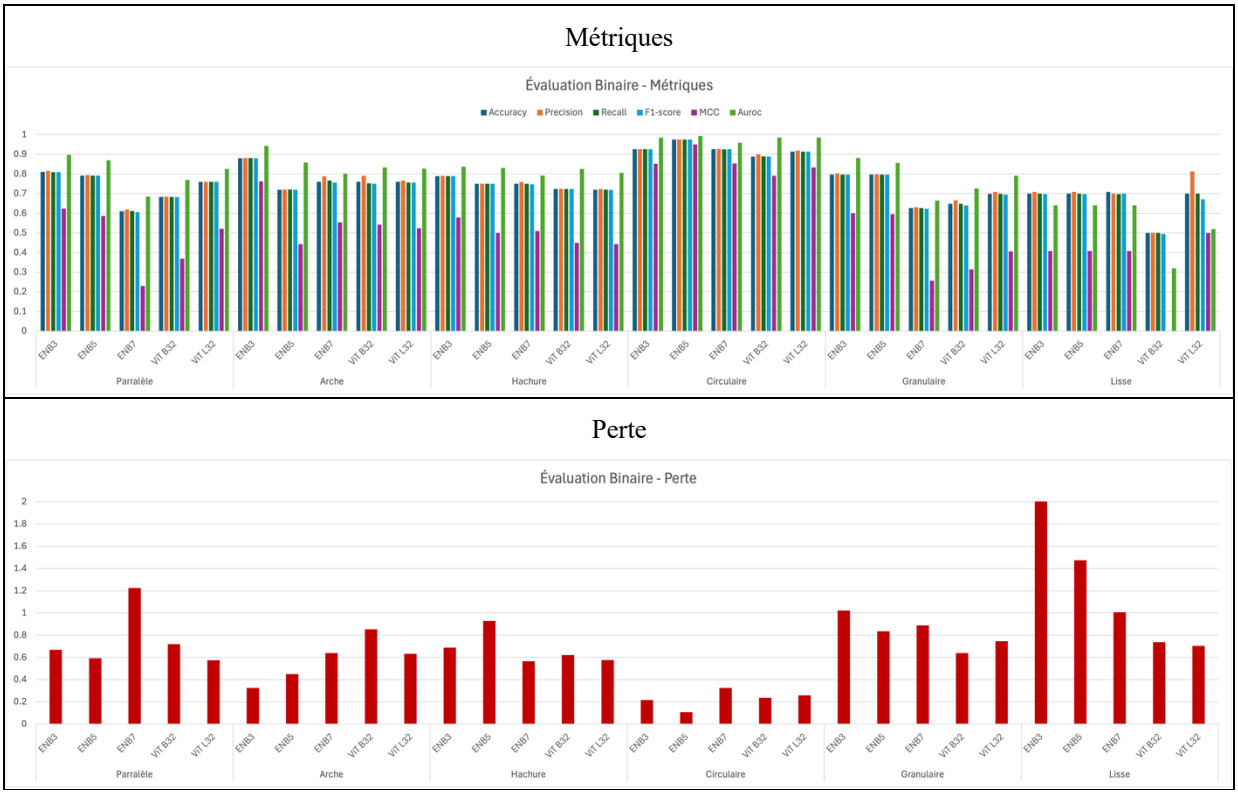


Figure 4.18 Graphiques comparatifs des métriques d'évaluation et de la perte pour tous les modèles binaires sans augmentation

Nous avons ensuite compilé les matrices de confusion binaire pour chacun des modèles. Les matrices pour les modèles ENB3, avec données augmentées et sans augmentation, sont présentées à la Figure 4.19. Tout comme pour les modèles multiétiquettes de la section précédente, nous remarquons les modèles avec le plus grand nombre d'erreurs de prédictions sont ceux pour les marques parallèle, hachure et granulaire. La Figure 4.20 présente les courbes ROC (Receiver Operating Curve) pour les deux modèles ENB3 (données augmentées et sans augmentation). Ces courbes, accompagnées de la mesure AUROC (l'aire sous la courbe), viennent appuyer ce que nous avons constaté lors de l'observation des métriques, c'est-à-dire que l'augmentation des données n'améliore pas les modèles pour les marques parallèle, arche, hachure et lisse. En comparant les mesures d'AUROC des modèles sans augmentation, nous confirmons que la performance du modèle pour la marque lisse est faible, celle des modèles pour les marques parallèle, hachure et granulaire est modérée et que celle des modèles pour les marques arche et circulaires est forte.

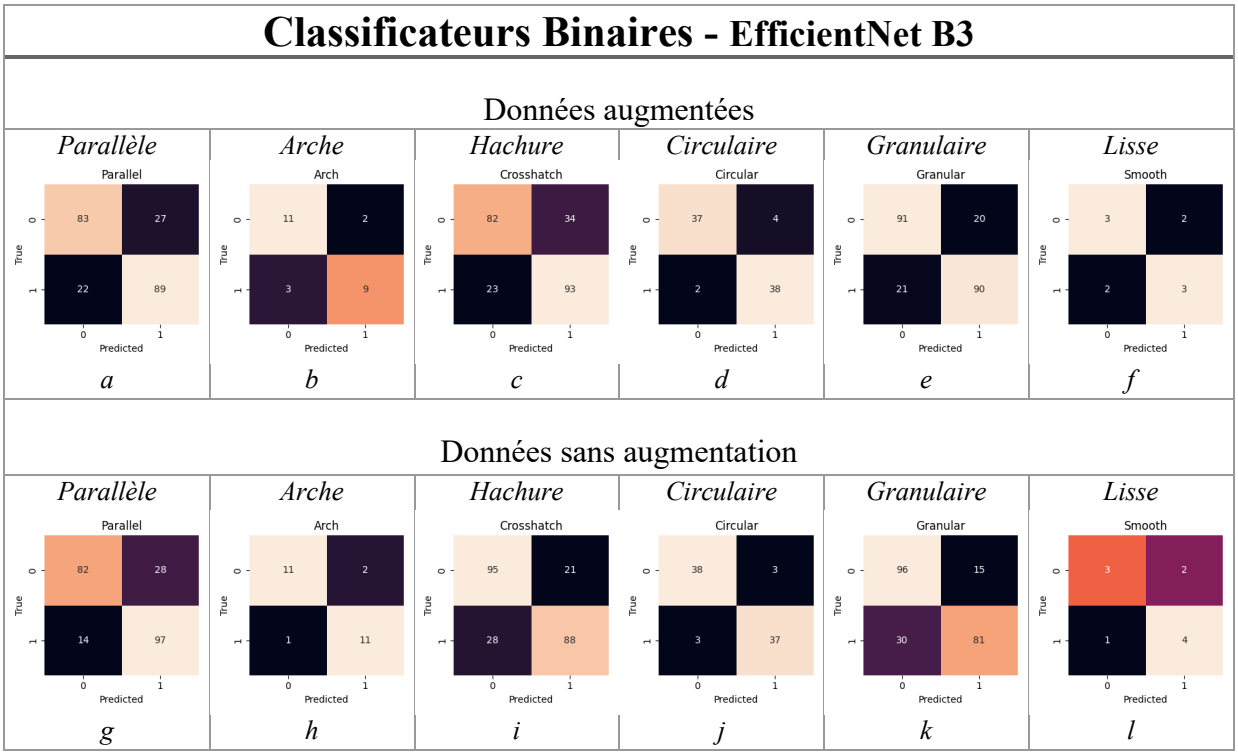


Figure 4.19 Matrices de confusion binaires pour les modèles ENB3 binaires : entraînés avec données augmentées (a-f) et sans augmentation (g-l)

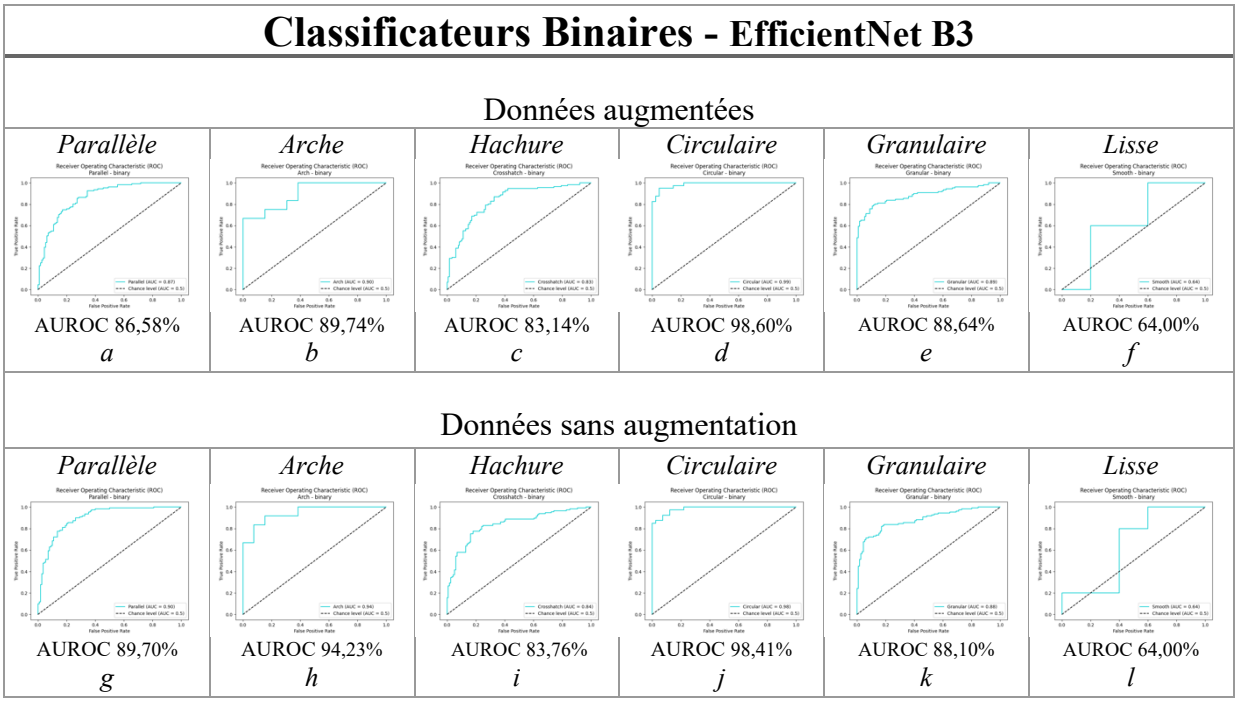


Figure 4.20 Courbes ROC pour les modèles ENB3 binaires : entraînés avec données augmentées (a-f) et sans augmentation (g-l)

Pour terminer, les cartes thermographiques des modèles binaires ENB3 (sans augmentation) sont présentées en annexe (ANNEXE III, Figure-A III-5). On peut observer que dans certains cas, les modèles ciblent bien la zone où se trouvent les marques recherchées, mais que d'autres fois, l'attention n'est pas du tout dans la région d'intérêt, même s'il arrive à faire la bonne prédiction.

4.3.4 Accord interannotateur

Afin de vérifier la validité des étiquettes, nous avons d'abord programmé une application en python, permettant de parcourir les images et d'annoter les classes observées. Puis, nous avons sélectionné 101 échantillons parmi toutes les images 2D provenant du HW3. Ces échantillons ont été choisis afin de représenter le plus grand nombre de catégories et de combinaisons de catégories possibles. Nous avons tenté de choisir des échantillons qui n'étaient ni trop faciles, ni trop complexes. Autrement dit, nous avons tenté de choisir les images les plus représentatives des catégories, sans être trop évidentes. Finalement, nous

avons recruté un second expert afin d'annoter les 101 échantillons choisis. Il faut préciser que cet étiquetage s'est fait à partir de l'image 2D uniquement, sans fonctionnalités de visualisation permettant de déplacer et de tourner les images ou de modifier l'éclairage. La possibilité d'examiner la vue 3D via la topographie était aussi absente de l'exercice. Nous avons ensuite comparé les étiquettes de l'annotateur original (annotateur A) avec les étiquettes de ce deuxième annotateur (annotateur B) en calculant le coefficient Kappa pour chaque catégorie. Le Tableau 4.13 présente les résultats obtenus. Le Tableau 4.14 présente le nombre d'étiquettes identifiées par chaque annotateur.

Tableau 4.13 Coefficient Kappa pour chaque catégorie

Catégorie	Kappa	Interprétation
Parallèle	0,57	Accord modéré
Arche	0,53	Accord modéré
Hachures	0,37	Accord léger
Circulaire	0,55	Accord modéré
Granulaire	0,02	Accord absent à faible
Lisse	0,06	Accord absent à faible
Inconnu	0,00	Absence d'accord
Moyenne	0,30	Accord léger

Tableau 4.14 Nombre d'étiquettes identifiées par chaque annotateur

Nb d'étiquettes	Annotateur A	Annotateur B
1	22	54
2	32	43
3	32	4
4	13	0
5	2	0

Nous observons que trois des catégories semblent plus évidentes à identifier (parallèle, arche et circulaire), car les annotateurs y obtiennent une mesure d'accord plus élevée. La catégorie

hachure semble plus difficile à identifier pour les annotateurs, car le coefficient Kappa chute à un accord léger. Les catégories granulaire et lisse sont presque au niveau du désaccord. Et il n'y a aucune entente pour les échantillons de catégorie inconnu. Pour le nombre d'étiquettes, nous remarquons que l'annotateur A utilise une plus grande variété dans le nombre d'étiquettes : les échantillons ont majoritairement deux ou trois étiquettes, mais il y a aussi un nombre important d'échantillons avec une ou quatre étiquettes. Deux échantillons ont été identifiés avec cinq étiquettes. Le second annotateur a surtout assigné une ou deux étiquettes aux échantillons. Il a considéré seulement quatre échantillons comme ayant trois étiquettes et aucun échantillon avec plus de trois étiquettes. Lors de cet exercice, nous avons aussi constaté qu'il est dommage que l'étiquetage multiétiquette ne tienne pas compte de la proportion de chaque catégorie dans l'image. Par exemple, une image étiquetée « parallèle, granulaire » pourrait avoir des marques parallèles sur 90% de sa surface et une section minimale avec des marques granulaires. Les mêmes étiquettes seront appliquées à une image contenant des marques granulaires sur 90% de la surface et une section minimale contenant des marques parallèles. Dans la réalité, ces deux images auraient une apparence bien différente. À la suite de ces observations, nous croyons que les modèles de classification sont limités par l'incohérence des étiquettes. Nous ne croyons pas réussir à obtenir des résultats supérieurs en recherchant des architectures mieux adaptées ou en réglant les hyperparamètres. Un travail de ré-étiquetage serait probablement nécessaire afin d'améliorer les résultats obtenus.

4.4 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle méthode pour l'identification des marques sur des images de douilles de cartouche avec des modèles de classification multiclasses, des modèles multiétiquettes, et des modèles binaires. Pour les modèles multiclasses, nous avons obtenu des résultats passables, avec le modèle ENB3, entraîné à partir des données HW4, augmentées et équilibrées : 77,85% pour l'exactitude, 71,53% pour la mesure F1, mais avec une valeur de perte élevée à 1,20. Ces résultats nous ont amenés à contester la validité de l'étiquetage et à mener une brève vérification. Les résultats nous ont

confirmé que l'accord interannotateur est modéré pour les marques parallèle, arche et circulaire (avec un coefficient Kappa de 0,57; 0,53 et 0,55; respectivement); léger pour la marque hachure (avec un coefficient Kappa de 0,37); et presque absent pour les marques granulaire et lisse (avec un coefficient Kappa de 0,02 et 0,06). L'accord est complètement absent pour la catégorie des inconnus (que nous avions déjà exclus des expérimentations).

Pour les modèles multiétiquettes, nous avons obtenu des résultats que nous jugeons insuffisants, avec le modèle ENB3, entraîné à partir des données HW4, sans augmentation : 52,88% pour l'exactitude, 54,47% pour la mesure F1, avec une valeur de perte de 0,36. Nous avons constaté que l'augmentation avec une stratégie d'équilibrage des données ne fonctionnait pas bien pour les ensembles multiétiquettes, lorsque les données sont déséquilibrées sur deux niveaux. Nous croyons que les modèles de classification sont limités par l'incohérence des étiquettes et que ces limitations sont encore plus évidentes dans les modèles multiétiquettes, où les prédictions sont plus complexes. Le ré-étiquetage des échantillons, ainsi que l'utilisation d'une fonction de perte tenant compte du déséquilibre, telle que la perte focale, pourrait probablement améliorer ces résultats.

Les modèles binaires entraînés à partir des données HW4, sans augmentation, sont ceux qui obtiennent les meilleurs résultats d'évaluation. Les modèles atteignent les valeurs suivantes (pour la mesure F1 et la perte) : parallèle (81% et 0,67); arche (88% et 0,33); hachure (79% et 0,69); circulaire (93% et 0,22); granulaire (80% et 1,02); et lisse (70% et 2,01). Pour tous ces modèles, la valeur de l'exactitude est presque identique à celle de la mesure F1. On remarque que les pertes les plus élevées sont pour les modèles granulaire et lisse, les deux catégories ayant obtenu un accord presque inexistant entre les annotateurs. Même si l'incohérence des étiquettes limite les modèles, les courbes d'entraînement et les résultats d'évaluation sont prometteurs lorsque la classification est effectuée séparément pour chaque type de marque. Nous croyons qu'il est plus facile d'équilibrer les ensembles de données lorsque l'on se concentre sur une classe à la fois. Malgré tout, nous croyons que les résultats pourraient être améliorés avec un ré-étiquetage des échantillons.

CHAPITRE 5

CONTRIBUTION #2 – APPRENTISSAGE NON SUPERVISÉ POUR LE REGROUPEMENT DE MARQUES MICROSCOPIQUES

5.1 Introduction

Ce chapitre porte sur la deuxième contribution qui vise à évaluer des regroupements d'images de douilles de cartouche, obtenus à l'aide de méthodes d'apprentissage non supervisé. Il débute avec une présentation de la méthodologie; incluant le choix des algorithmes et des métriques, la préparation des données, ainsi que le détail de l'implémentation. Le chapitre se poursuit avec une présentation des résultats pour les différentes expérimentations, accompagnée d'observations.

5.2 Méthodes

Les marques de classe sur les douilles de cartouche sont difficiles à distinguer, et puisque le problème est multiétiquette, plusieurs marques pourraient se chevaucher. Les expérimentations de ce chapitre sont divisées en deux thèmes principaux: le regroupement par algorithme de clustering et le regroupement par réseau profond de clustering.

5.2.1 Algorithmes

Pour le regroupement par algorithme de clustering, nous avons sélectionné les algorithmes K-means, Fuzzy C-means, DBSCAN, HDBSCAN, OPTICS et spectral. Nous avons d'abord choisi K-Means parce qu'il s'agit d'un algorithme simple, facile à comprendre et rapide à exécuter. Ces qualités en font un choix populaire pour l'exploration non supervisée d'un ensemble de données et l'analyse de ses groupes. K-Means effectue le regroupement en se basant sur un partitionnement, et débute en déterminant aléatoirement les positions d'un nombre fixe (k) de centroïdes. Ensuite, lors de chaque itération, les données sont assignées aux centroïdes les plus proches et les positions des centroïdes sont recalculées en utilisant la

moyenne des groupes. Le principal inconvénient de cette technique est sa performance, souvent inférieure à celle des algorithmes plus complexes, puisque des variations légères dans les données peuvent provoquer une variance élevée. On lui reproche aussi de créer des groupes sphériques de tailles égales, ce qui peut réduire la performance si les données ne correspondent pas à cette distribution (Education Ecosystem, 2018). L'algorithme Fuzzy C-means, dont le fonctionnement est similaire à celui de K-means, se décrit comme un clustering flou. C'est-à-dire que les échantillons situés aux frontières des clusters peuvent appartenir à plus d'une catégorie. En plus de fournir une liste de l'étiquette prédite reliant chaque échantillon à un cluster, l'algorithme Fuzzy C-means calcule une matrice de partition floue, qui indique le degré d'appartenance de chaque échantillon à chaque cluster. Ces algorithmes, basés sur le centroïde, sont des algorithmes de type inductif, c'est-à-dire qu'ils peuvent être appliqués directement sur de nouvelles données sans devoir recalculer les clusters.

Les algorithmes transductifs, quant à eux, doivent recalculer les clusters avant de faire des prédictions sur un nouvel échantillon. Parmi ceux-ci se trouvent DBSCAN, HDBSCAN et OPTICS, trois approches basées sur la densité. C'est-à-dire que l'algorithme recherche des échantillons noyau *(core samples)* appartenant à des régions de densité élevée et élargit les groupes à partir de ces noyaux. Contrairement à K-Means, les groupes peuvent être de formes variées et de tailles différentes. DBSCAN fonctionne bien lorsque la densité des clusters est homogène globalement, alors que HDBSCAN explore toutes les échelles de densité. OPTICS utilise un critère additionnel de distance d'accessibilité, permettant de former des groupes de densité variable. Pour les trois algorithmes, le nombre de clusters n'est pas déterminé au départ : il dépend plutôt de certains paramètres, tels que la distance maximum du centre dense ou le nombre minimum de points dans un groupe. De plus, ces algorithmes ne classent pas tous les points dans les clusters : plusieurs se retrouvent souvent à l'extérieur des clusters. Ils sont alors identifiés comme étant des valeurs aberrantes ou du bruit. Les approches basées sur la densité sont plus efficaces avec les données dont la distribution des groupes n'est pas régulière, ou avec des données qui contiennent beaucoup de bruit ou de valeurs aberrantes, mais elles sont moins efficaces lorsque les données sont

disposées en groupes plus normaux (Yıldırım, 2020). Le dernier algorithme transductif étudié est le clustering spectral, une approche basée sur les distances entre les graphes. De façon générale, le clustering spectral est reconnu pour se démarquer lorsque la structure des clusters n'est pas de forme convexe. En d'autres termes, lorsqu'on ne peut pas décrire les clusters par des mesures de centre accompagnées de mesures de dispersion.

Pour le regroupement par réseau profond de clustering, nous avons sélectionné le DCN proposé par (Yang *et al.*, 2017), une architecture basée sur les autoencodeurs. Dans cet article, les auteurs proposent une architecture optimisant simultanément la réduction de la dimensionnalité et le clustering K-means, afin de trouver un espace latent favorisant la formation des clusters. Le réseau DCN se compose de trois parties : un encodeur, un décodeur et un module de clustering. Le rôle de l'encodeur consiste à convertir l'entrée de dimensions élevées en un espace latent de dimensions réduites. Le décodeur permet de s'assurer que l'entrée peut être reconstruite à partir de cet espace latent, et permet ainsi d'éliminer les solutions triviales. Finalement, le module de clustering pénalise les solutions de l'espace latent qui sont moins propices à créer des clusters K-means. La Figure 5.1 montre le diagramme de l'architecture DCN, tel que présenté par les auteurs.

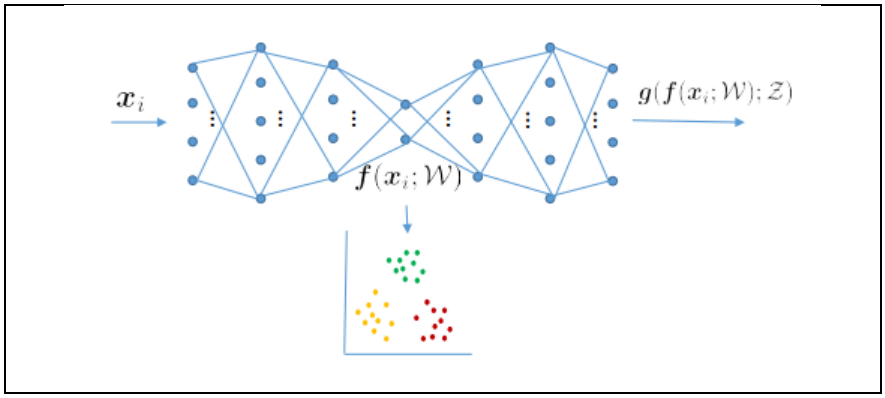


Figure 5.1 Diagramme du réseau profond de clustering (DCN)

$$\sum_{i=1}^{N}\left(\ell(g(f(x_i)),x_i)+\frac{\lambda}{2}\|f(x_i)-Ms_i\|_2^2\right) \tag{5.1}$$

$$\ell(x,y)=\|x-y\|_2^2 \tag{5.2}$$

L'équation (5.1 présente la fonction de perte combinée, utilisée par le réseau. Le premier terme, $\ell(g(f(x_i)),x_i)$, représente l'erreur de reconstruction; avec $f(x_i)$ la fonction de mappage de l'encodeur, et $g(f(x_i))$ la fonction de reconstruction du décodeur à partir de l'espace latent. Le symbole ℓ représente une fonction de perte pour la reconstruction. Le DCN utilise la fonction de perte des moindres carrés (least square loss), présentée à l'équation (5.2), mais les auteurs précisent que d'autres fonctions de perte peuvent être utilisées, telles que la norme L1 ou la divergence KL. Le second terme, $\frac{\lambda}{2}\|f(x_i)-Ms_i\|_2^2$, représente l'erreur de clustering. Le symbole λ est un paramètre de régularisation permettant d'équilibrer l'erreur de clustering par rapport à l'erreur de reconstruction afin d'améliorer la découverte d'une représentation latente compatible avec K-means. Au départ, nous avions aussi envisagé d'autres options de clustering profond telles que le clustering intégré profond (DEC), aussi basé sur les autoencodeurs et le réseau adversarial génératif maximisant l'information (InfoGAN), basé sur les modèles génératifs (Dahal, 2019), mais ces expérimentations n'ont pas été effectuées. Elles pourraient faire partie de travaux futurs.

5.2.2 Métriques

Pour l'évaluation des clusters, nous avons d'abord utilisé le coefficient Silhouette. Cette métrique permet d'évaluer la qualité technique des clusters et ne nécessite pas de connaître la vérité de terrain. Nous avons ainsi pu l'utiliser avec toutes les expérimentations. Cette métrique est calculée pour chaque échantillon et un graphique tracé à partir de ces valeurs permet d'effectuer une évaluation visuelle. De plus, le coefficient Silhouette moyen présente une vue d'ensemble permettant la comparaison. Nous avons aussi utilisé des graphiques en

nuages de points afin de représenter les groupes formés. Ces informations nous permettent d'observer rapidement le comportement des algorithmes avec nos données.

Pour les expérimentations avec les données étiquetées, la vérité de terrain était disponible et nous avons utilisé les métriques suivantes : l'indice de Rand ajusté (ARI), l'information mutuelle ajustée (AMI), l'indice Fowlkes-Mallows (FM), la mesure V (VM), l'homogénéité (H) et la complétude (C). De façon générale, la métrique ARI est préférable lorsque la taille des clusters est large et uniforme, alors que la métrique AMI est préférable lorsque les données sont débalancées et qu'il est possible de retrouver des clusters de petite taille. La métrique FM s'intéresse à la qualité des clusters, en calculant un rapport entre la précision et le rappel, alors que les métriques d'homogénéité, de complétude ainsi que la Mesure V sont comparables aux métriques de précision, de rappel et de mesure F1, utilisées pour l'évaluation de la classification supervisée. Il est tout de même opportun de rappeler que la vérité de terrain a été identifiée comme étant incertaine dans le chapitre précédent. Les métriques utilisant cette vérité de terrain sont donc à considérer avec un certain détachement.

5.2.3 Préparation des données

Pour les expérimentations de clustering, nous avons débuté en utilisant les 3 609 images étiquetées provenant du BrassTRAX HW4 (section 3.3.2). Pour chaque algorithme, nous avons reproduit quelques expérimentations avec les données supplémentaires non étiquetées (7 402 échantillons provenant de deux autres appareils BrassTRAX HW4), en nous basant sur les paramètres préalablement optimisés sur les données étiquetées. Pour la suite de ce chapitre, ces deux ensembles seront nommés **« données étiquetées »** et **« données non étiquetées »**, respectivement. Pour chacun des algorithmes, à moins d'une spécification du contraire, les expérimentations sont effectuées à partir des **données étiquetées**.

Nous avons utilisé les images avec la configuration d'entrée choisie pour les expérimentations de l'objectif précédent. Ces images topographiques 3D sont filtrées et fusionnées avec le masque de la région d'intérêt remplacé par du bruit gaussien aléatoire

(Figure 4.1 e et j). Les algorithmes de clustering, ainsi que le réseau profond de clustering, ont été appliqués sur les ensembles de données sans division préalable en sous-ensembles ni augmentation. Nous avons mis les étiquettes de côté pour faire les regroupements, puis nous les avons utilisées pour calculer certaines métriques. Nous avons aussi inclus les images appartenant à la catégorie « inconnu », qui avaient été exclues des expérimentations du chapitre précédent. Nous croyons que ces échantillons seront regroupés avec des clusters pertinents, ce qui pourrait ajouter de l'incertitude aux métriques utilisant la vérité de terrain.

5.2.4 Implémentation

Avant d'appliquer un algorithme de clustering sur des images, il faut d'abord en extraire les caractéristiques descriptives. Pour ce faire, nous avons utilisé les modèles de CNN et de ViT, préentraînés pour la classification sur ces mêmes données lors de l'objectif précédent. Nous avons supprimé la tête de classification des modèles, afin d'obtenir une sortie correspondant à un vecteur de caractéristiques descriptives de l'image en entrée. Pour chacun des trois modèles choisis (VGG16, ENB3 et ViT), l'avant-dernière couche a été utilisée afin d'extraire les caractéristiques. Dans les trois cas, nous avons obtenu un vecteur de 32 caractéristiques descriptives. Nous avons aussi utilisé un autoencodeur, que nous avons entraîné sur nos données. L'espace latent correspondant à la sortie de l'encodeur fournit un vecteur de 64 caractéristiques descriptives. Les détails de l'autoencodeur sont présentés plus bas.

Une fois les vecteurs de caractéristiques descriptives obtenus, nous les avons mis à l'échelle et normalisés. Puis, nous avons utilisé un algorithme de réduction de la dimensionnalité. Nous avons testé deux algorithmes populaires : l'analyse en composantes principales (PCA), provenant de la librairie de décomposition de *sklearn* (Scikit-learn, [s d]); et l'intégration de voisins stochastiques distribués (TSNE), provenant de la librairie manifold de *sklearn* (Scikit-learn, [s d]). Finalement, nous avons appliqué les algorithmes de clustering sur les caractéristiques descriptives réduites. Nous avons utilisé les algorithmes de clustering K-means, DBSCAN, HDBSCAN, OPTICS et Spectral provenant de la librairie de clustering de *sklearn* (Scikit-learn, [s d]). Nous avons utilisé l'algorithme Fuzzy C-means disponible sur

Pypi (Dias, [s d]). L'implémentation pour les métriques provient de la librairie de métriques de *sklearn* (Scikit-learn, [s d]). Le diagramme de la Figure 5.2 illustre le processus général de clustering.

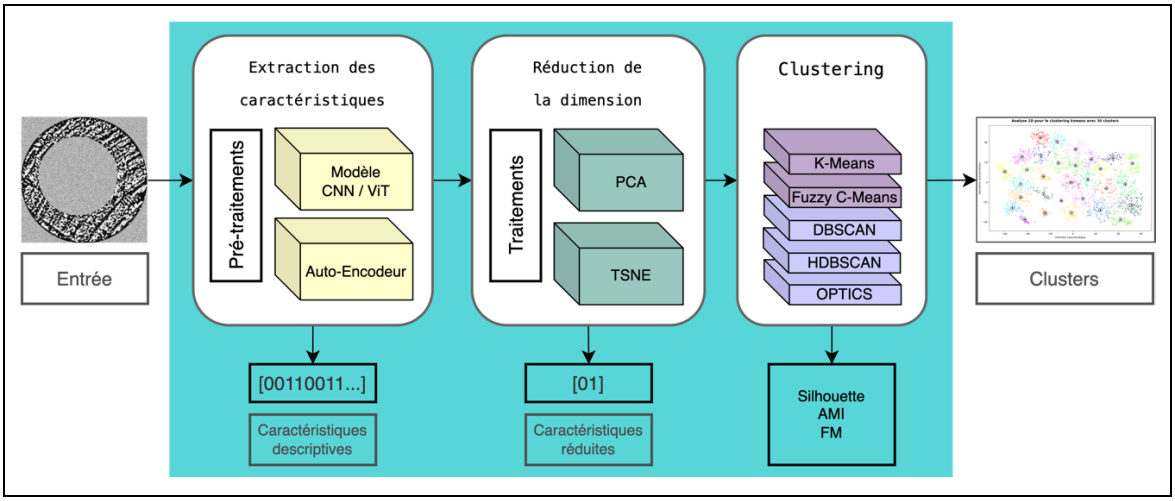


Figure 5.2 Diagramme du processus de clustering

5.2.4.1 Autoencodeurs

Afin d'évaluer les autoencodeurs comme méthode d'extraction des caractéristiques, nous avons implémenté un autoencodeur simple, et nous l'avons entraîné pour la reconstruction des images à partir des données étiquetées (AE-I). Nous avons utilisé une fonction d'activation gelu pour les couches denses, et une fonction d'activation sigmoïde en sortie. Nous avons utilisé la perte d'entropie croisée binaire et effectué l'entraînement sur dix époch. Les métriques d'entraînement étaient faibles lorsque nous avons utilisé les images contenant du bruit gaussien, tel que dans les expérimentations précédentes; alors nous avons entraîné un second autoencodeur à partir des images sans bruit gaussien (AE-II). La Figure 5.3 présente un diagramme de l'autoencodeur.

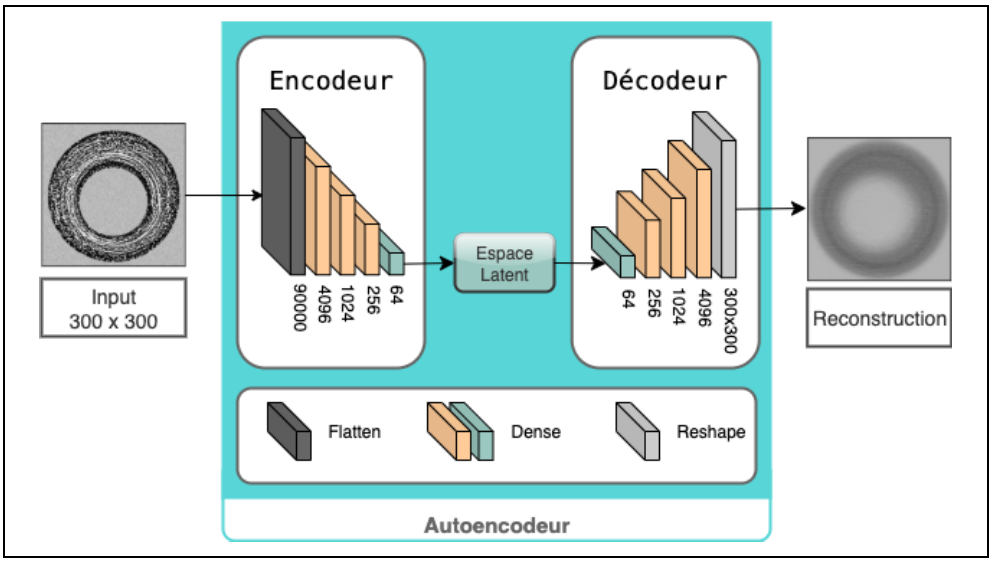


Figure 5.3 Diagramme de l'autoencodeur

5.2.4.2 Clustering spectral

Pour un survol de la technique, le clustering spectral commence par transformer les données en un graphique, à partir duquel il crée une matrice de similarité. Cette matrice, carrée et symétrique, représente les distances entre des paires d'échantillons (pairwise). Avec l'implémentation de l'algorithme que nous avons utilisé, un paramètre permet de sélectionner différents noyaux afin de construire cette matrice de similarité. Par exemple le noyau RBF (radial basis function) ou le noyau Laplacien. De plus, selon le noyau choisi, différents paramètres peuvent être configurés. Par exemple *gamma* pour les noyaux RBF, poly, sigmoid, Laplacian et chi2. L'algorithme calcule ensuite le spectre de cette matrice de similarité; c'est-à-dire les vecteurs propres (eigenvectors) et les valeurs propres (eigenvalues), triés en ordre d'importance (les vecteurs propres les moins significatifs étant ceux possédant les valeurs propres les moins élevées). Pour cette étape, un paramètre *eigen_solver* permet de sélectionner la stratégie de décomposition des valeurs propres, et trois options sont disponibles : arpack, lobpcg et amg. Les *n* vecteurs propres les moins significatifs sont ensuite utilisés afin de représenter chaque échantillon en *n* dimensions. Cette étape représente une opération de réduction de la dimensionnalité non linéaire. Un paramètre *n_components* permet de spécifier le nombre de vecteurs propres à utiliser pour les

plongements vectoriels. Finalement, un algorithme de clustering est appliqué sur ces *n* vecteurs propres. La sélection de cet algorithme se fait par le paramètre *assign_labels* et trois options sont possibles : kmeans, discretize et cluster_qr. Il faut aussi configurer le paramètre *n_clusters* afin d'indiquer le nombre de clusters recherchés.

5.2.4.3 Réseau profond de clustering

Pour les expérimentations de regroupement par réseau profond de clustering, nous avons omis les étapes d'extraction des caractéristiques et de réduction de la dimensionnalité, car ces étapes sont prises en charge par le DCN. Nous avons utilisé une implémentation du DCN disponible sur GitHub (Morris, 2021), qui reprend l'idée de l'auteur original pour offrir une solution utilisant Tensorflow et Keras; et optimisée sur Python 3.9. Outre les paramètres configurables, nous avons utilisé le DCN tel qu'implémenté sur GitHub, sans modifier le type de réseau profond, la méthode de clustering ou les fonctions de perte utilisés. Nous avons remplacé la fonction *tf.norm*, utilisée dans les calculs des pertes, par les fonctions *tf.sqrt*, *tf.reduce_sum* et *tf.pow*. Cette correction visait à régler un problème rencontré avec la fonction *tf.norm* qui retournait des valeurs NAN plutôt que des valeurs de perte lors de l'entraînement, nuisant ainsi au bon fonctionnement du réseau.

5.3 Résultats

5.3.1 Entraînement des autoencodeurs

La Figure 5.4 présente les courbes de performance et d'optimisation obtenues lors de l'entraînement des deux autoencodeurs. Il est à noter que les graphiques des deux autoencodeurs sont sur des échelles différentes, car les différences de valeur étaient trop élevées pour pouvoir utiliser la même échelle sans perdre le détail des courbes. La Figure 5.5 montre des exemples des images reconstituées par chacun des deux modèles. Une fois l'entraînement des autoencodeurs terminé, nous avons utilisé l'espace latent à la sortie de l'encodeur pour extraire les caractéristiques des images.

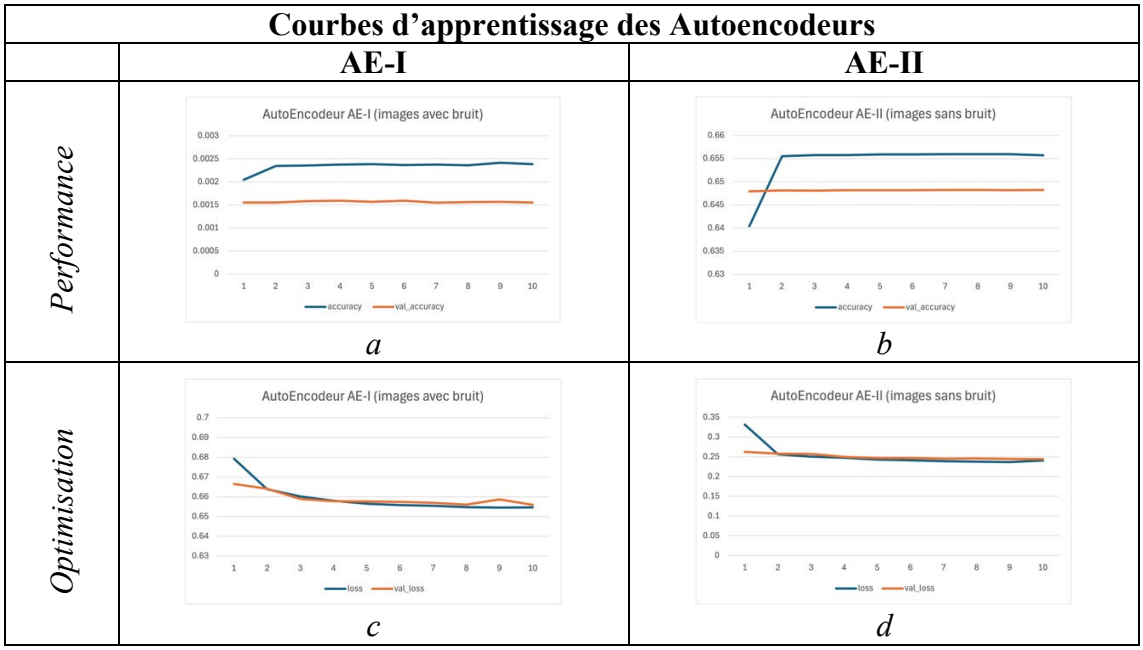


Figure 5.4 Courbes d'apprentissage des autoencodeurs entraînés avec les données étiquetées. AE-I : entraînement à partir des images avec bruit gaussien (a et c) et AE-II : entraînement à partir des images sans bruit (b et d)

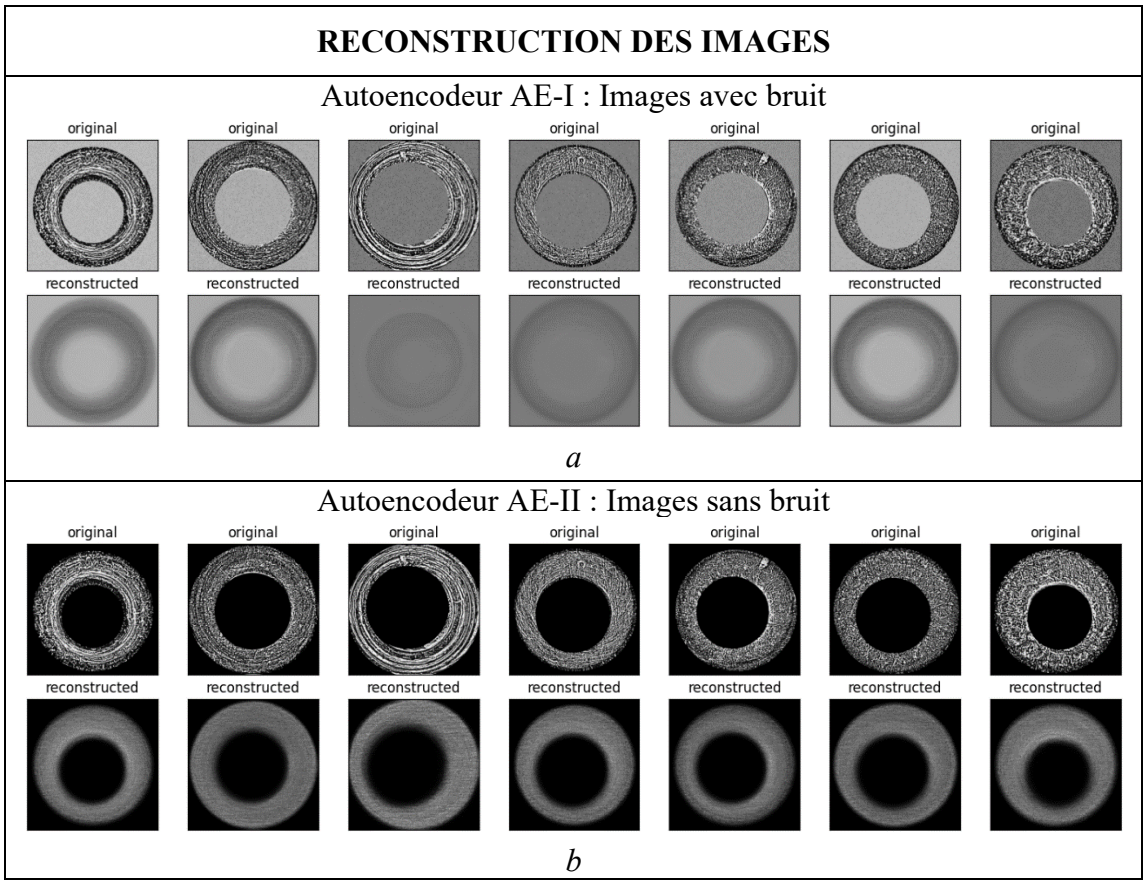


Figure 5.5 Images reconstruites des autoencodeurs entraînés avec les données étiquetées. AE-I : Images avec bruit gaussien (a) et AE-II : Images sans bruit (b)

5.3.2 Regroupement par algorithme de clustering

5.3.2.1 K-means

Nous avons débuté les expérimentations avec l'algorithme de clustering K-means, appliqué sur les caractéristiques extraites par le modèle ENB3 préentraîné. Pour réduire les dimensions des caractéristiques, nous avons testé les algorithmes PCA et TSNE avec 2 et 3 dimensions.

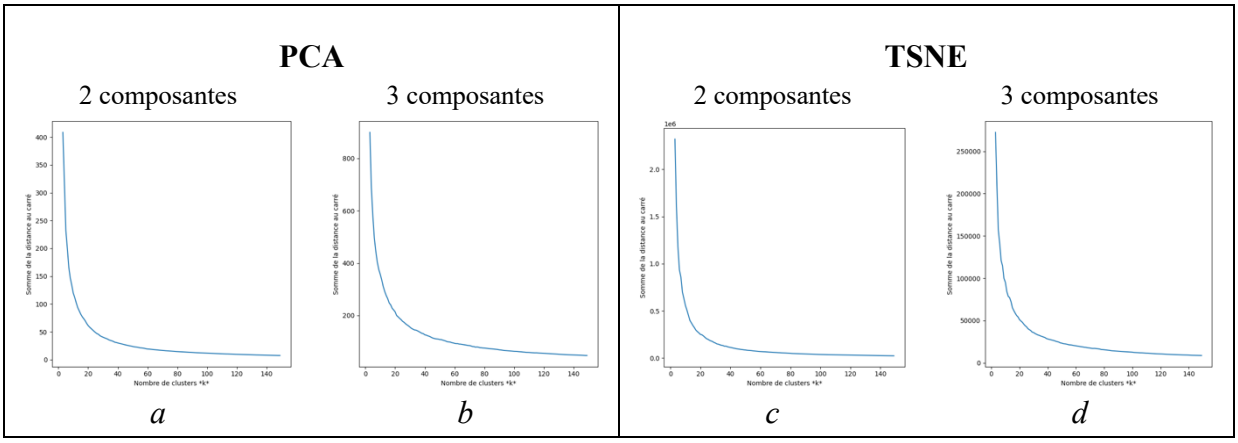


Figure 5.6 K-means : Courbes de la somme de la distance au carré selon le nombre de clusters. PCA avec 2 (a) et 3 composantes (b). TSNE avec 2 (c) et 3 composantes (d)

Avant d'appliquer l'algorithme K-means, il faut spécifier le nombre de clusters désiré. Ce paramètre est difficile à établir, en particulier lorsque les données ne sont pas étiquetées, car nous ne savons pas combien de groupes pourraient être présents. Même lorsque les étiquettes sont connues, cela ne veut pas dire qu'elles sont correctes, ou qu'elles correspondent à ce qui sera vu par l'algorithme. Afin de nous aider à déterminer le nombre de clusters, nous avons généré des courbes permettant de visualiser la somme de la distance au carré selon le nombre de clusters générés (Figure 5.6). Sur ces courbes, nous observons que la somme de la distance au carré semble atteindre un plateau autour de 40 clusters. Nous avons donc appliqué l'algorithme K-means avec un nombre de 40 clusters sur les résultats des quatre options de réduction de la dimensionnalité proposées : PCA 2 et 3 composantes, et TSNE 2 et 3 composantes. Les paramètres utilisés pour l'algorithme K-means sont présentés dans le Tableau 5.1 et le graphique de la Figure 5.7 présente les différentes métriques calculées. Nous pouvons observer que toutes les métriques sont supérieures lorsque les caractéristiques sont réduites à 2 dimensions par TSNE. Nous avons donc choisi d'utiliser cette technique de réduction pour la suite des expérimentations.

Tableau 5.1 Paramètres de l'algorithme K-means

Paramètre	**Valeur**
init	k-means++
n_init	auto
algorithm	lloyd
Random_state	27

Nous avons ensuite évalué l'algorithme K-means avec différentes valeurs pour le nombre de clusters. Les valeurs se situent entre 10 et 50, par incrément de 5. Le graphique de la Figure 5.8 présente les différentes métriques calculées. Nous pouvons observer que certaines métriques augmentent lorsque le nombre de clusters augmente (VM et homogénéité), alors que d'autres diminuent (ARI, FM et complétude). Ce qui nous indique qu'un compromis est nécessaire pour le nombre de clusters. En observant les métriques de Silhouette et AMI, nous constatons qu'un nombre de 30 clusters semble optimal. Nous avons effectué la majorité de nos expérimentations avec l'algorithme Lloyd, mais nous avons aussi testé l'algorithme Elkan. Puisque les deux algorithmes produisent des résultats semblables, nous n'avons pas inclus les résultats produits avec l'algorithme Elkan dans cette thèse.

Puis, nous avons généré les graphiques Silhouette ainsi que les graphiques en nuages de points pour un clustering K-means de 30 groupes, à partir des réductions à 2 dimensions (Figure 5.9) et de celles à 3 dimensions (Figure 5.10). Nous observons que les graphiques en nuages de points 3D pour PCA et TSNE à 3 composantes (Figure 5.10 b et d) semblent afficher des groupes moins bien séparés, qui se chevauchent par endroit. Cette observation confirme que la réduction à 3 dimensions réduit la qualité des clusters. En comparant les graphiques en nuages de points 2D pour PCA et TSNE à 2 composantes (Figure 5.9 b et d), nous remarquons que les formes des groupes ainsi que la forme générale des données sont différentes, et nous remarquons que les groupes semblent un peu plus espacés dans le graphique pour la réduction par TSNE.

K-means – 40 clusters

Données étiquetées. Extraction des caractéristiques par ENB3

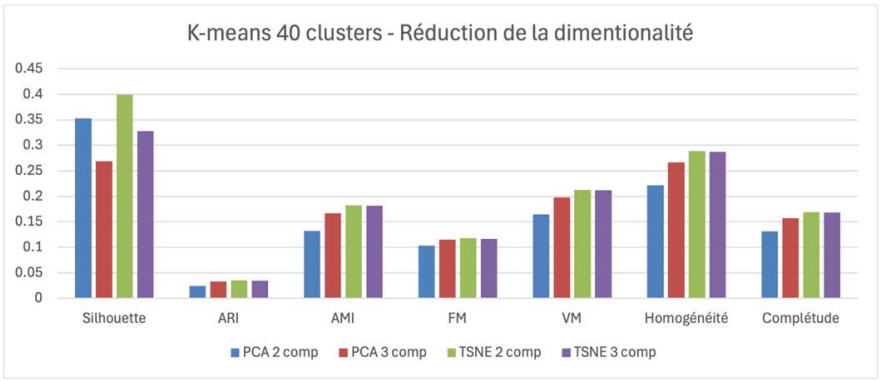


Figure 5.7 Graphique en barres des métriques d'évaluation des clusters K-means, avec différentes méthodes de réduction de la dimensionnalité

K-means

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

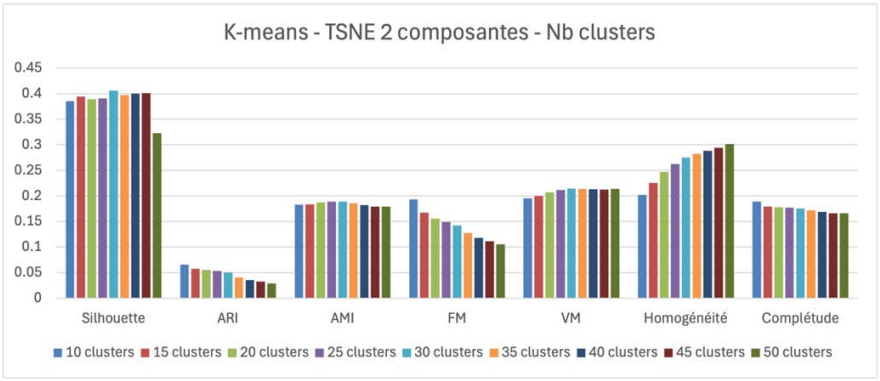


Figure 5.8 Graphique en barres des métriques d'évaluation des clusters K-means, avec variation du nombre de clusters

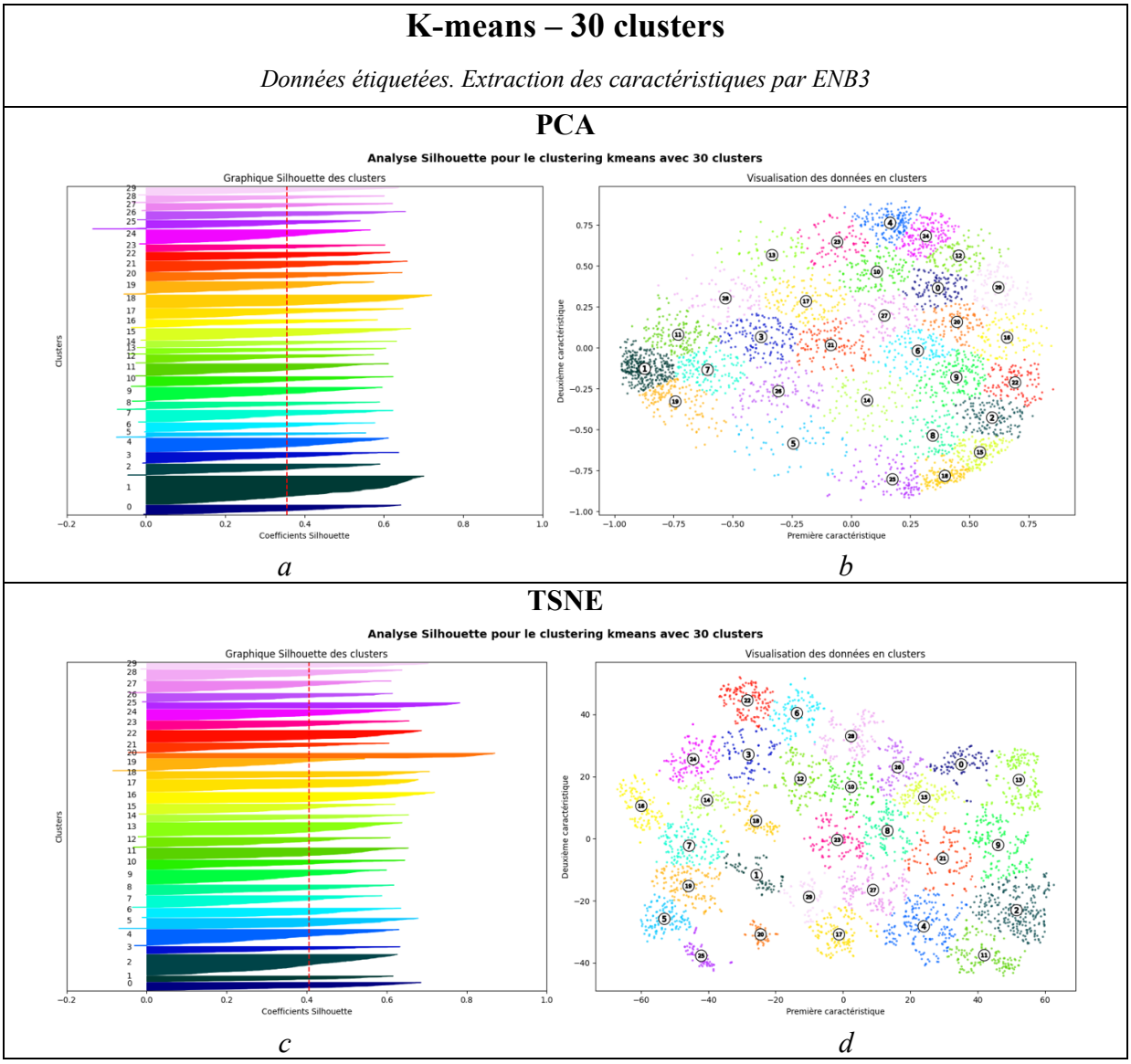
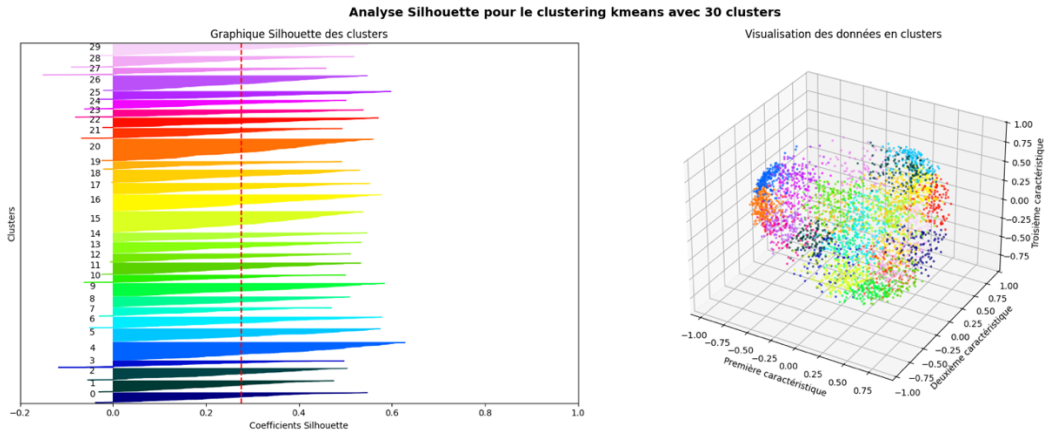


Figure 5.9 Clustering K-means. Réduction PCA en 2 composantes : graphique Silhouette (a), graphique 2D en nuages de points (b). Réduction TSNE en 2 composantes : graphique Silhouette (c), graphique 2D en nuages de points (d)

K-means – 30 clusters

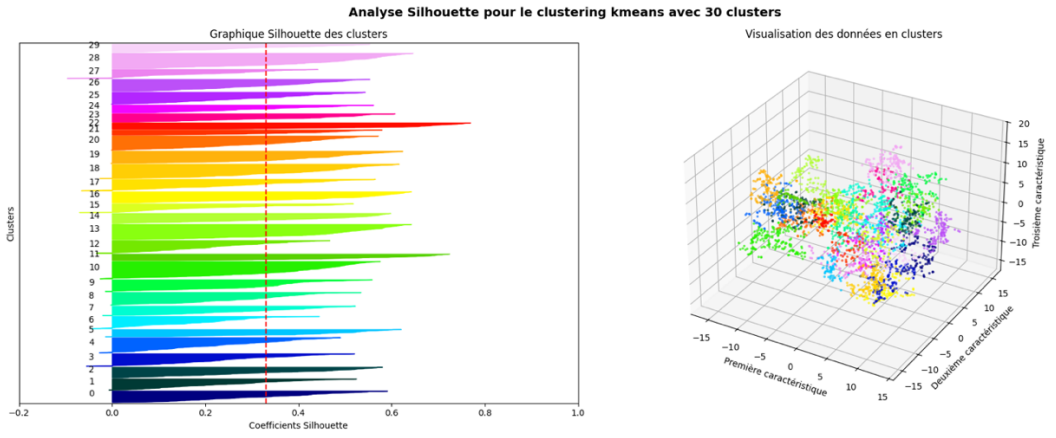
Données étiquetées. Extraction des caractéristiques par ENB3

PCA



a *b*

TSNE



c *d*

Figure 5.10 Clustering K-means. Réduction PCA en 3 composantes : graphique Silhouette (a), graphique 3D en nuages de points (b). Réduction TSNE en 3 composantes : graphique Silhouette (c), graphique 3D en nuages de points (d)

De plus, en comparant les graphiques Silhouette (Figure 5.9 a et c) nous observons que les groupes sont de meilleures qualités pour la réduction par TSNE, avec des valeurs de coefficient Silhouette plus élevées pour les échantillons. Nous observons que les clusters sont de taille uniforme, conformément à ce qui est attendu de l'algorithme K-means. Ce qui nous laisse croire que certains groupes ne sont pas optimaux, puisque certaines catégories devraient être plus rares que d'autres, et ces clusters devraient contenir moins d'échantillons.

Ensuite, nous avons évalué les différents modèles proposés pour l'extraction des caractéristiques des images. Pour chacune des trois architectures à l'étude: VGG16, ENB3 et ViTB32, nous avons utilisé le meilleur modèle préentraîné lors de l'objectif précédent. Nous avons aussi utilisé les deux autoencodeurs présentés précédemment afin d'évaluer si cette méthode pourrait être préférable.

K-means – 30 clusters

Données étiquetées. Réduction à 2 dimensions par TSNE

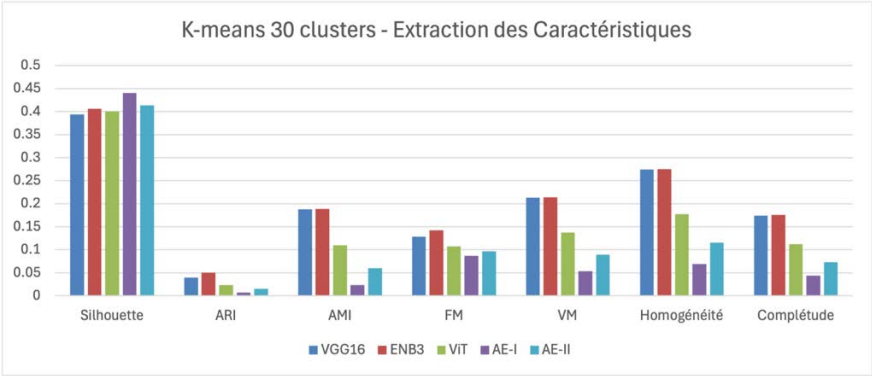


Figure 5.11 Graphique en barres des métriques d'évaluation des clusters K-means, avec différentes méthodes d'extraction des caractéristiques

Nous avons appliqué une réduction à 2 dimensions par TSNE sur les caractéristiques extraites à l'aide de chacune des méthodes évaluées, puis nous avons appliqué un clustering K-means de 30 clusters. Les métriques obtenues lors des évaluations sont présentées dans le graphique de la Figure 5.11. Nous observons que les coefficients Silhouette moyens sont plus élevés lorsque les autoencodeurs sont utilisés pour extraire les caractéristiques; alors que les autres métriques sont plus élevées lorsque les CNN sont utilisées.

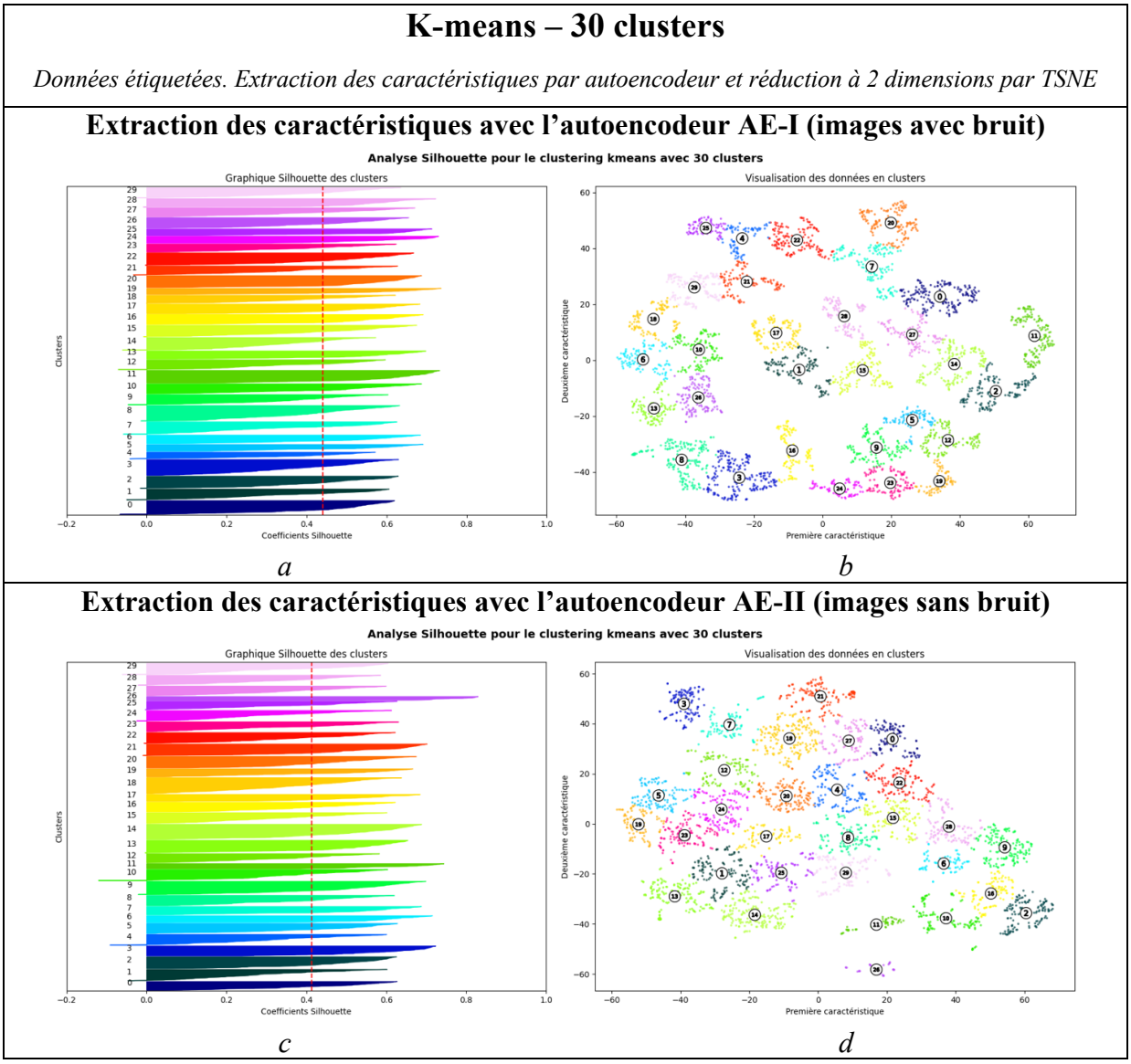


Figure 5.12 Clustering K-means. Extraction des caractéristiques par l'encodeur de AE-I : graphique Silhouette (a) et graphique 2D en nuages de points (b). Extraction des caractéristiques par l'encodeur de AE-II : graphique Silhouette (a) et graphique 2D en nuages de points (b)

Afin de comparer avec les résultats obtenus précédemment, nous avons généré les graphiques Silhouette, ainsi que les graphiques en nuages de points pour le clustering K-means à partir des caractéristiques extraites par les encodeurs des deux autoencodeurs (Figure 5.12). Dans les graphiques en nuages de points, nous remarquons que les clusters formés à partir des caractéristiques extraites par AE-I sont séparés davantage les uns des autres, expliquant le

coefficient Silhouette moyen plus élevé. Ce qui nous porte à croire que les caractéristiques apprises par cette méthode pourraient être plus discriminantes.

En observant les images 2D des échantillons présents dans les clusters, nous remarquons qu'en général, les images d'un même cluster possèdent des marques similaires. L'avis d'un expert serait nécessaire pour identifier les échantillons qui sont attribués aux mauvais groupes. Cependant, nous remarquons que plusieurs groupes possèdent quelques échantillons qui ne semblent pas afficher les mêmes types de marques. Deux exemples sont présentés à la Figure 5.13. Les clusters formés à partir des caractéristiques réduites provenant de l'AE-II (sans bruit) semblent fortement influencés par les teintes de gris et les formes générales de la douille. Autrement, nous ne remarquons pas de différence significative entre les trois méthodes d'extraction, en ce qui concerne la qualité des groupes formés.

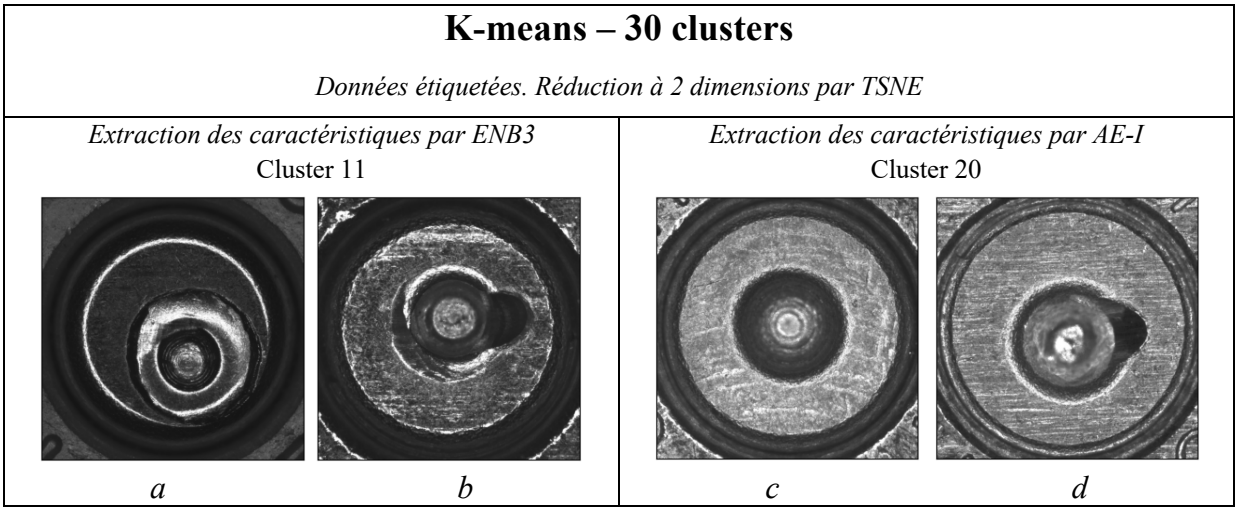


Figure 5.13 Échantillons provenant d'un même cluster et présentant des marques différentes : cluster 11 (a, b); cluster 20 (c, d)

Sans l'avis d'expert, nous ne pouvons que nous baser sur les métriques obtenues (Figure 5.11). Nous avons ainsi choisi de nous limiter à la méthode d'extraction par ENB3 pour le reste des expérimentations de ce chapitre. Cependant, nous croyons que des travaux futurs pourraient inclure l'exploration de modèles d'autoencodeur plus avancés permettant d'obtenir de meilleurs résultats, et ainsi permettre d'extraire des caractéristiques

significatives plus pertinentes. En résumé, pour l'algorithme K-means, nous avons choisi d'utiliser une extraction des caractéristiques par la CNN ENB3, une réduction à 2 dimensions par TSNE et un nombre de 30 clusters. Nous avons affiché les images 2D de cinq échantillons appartenant à chacun de ces 30 groupes en annexe (ANNEXE IV, Figure-A IV-1). À titre de comparaison, les images de l'expérimentation à partir des caractéristiques obtenues par la sortie de l'encodeur de l'autoencodeur AE-II sont présentées dans la Figure-A IV-2.

Finalement, pour terminer les expérimentations avec K-means, nous avons repris le clustering sur les **données non étiquetées**, avec ces mêmes paramètres. La Figure 5.14 montre le graphique Silhouette et le graphique en nuages de point pour cette expérimentation.

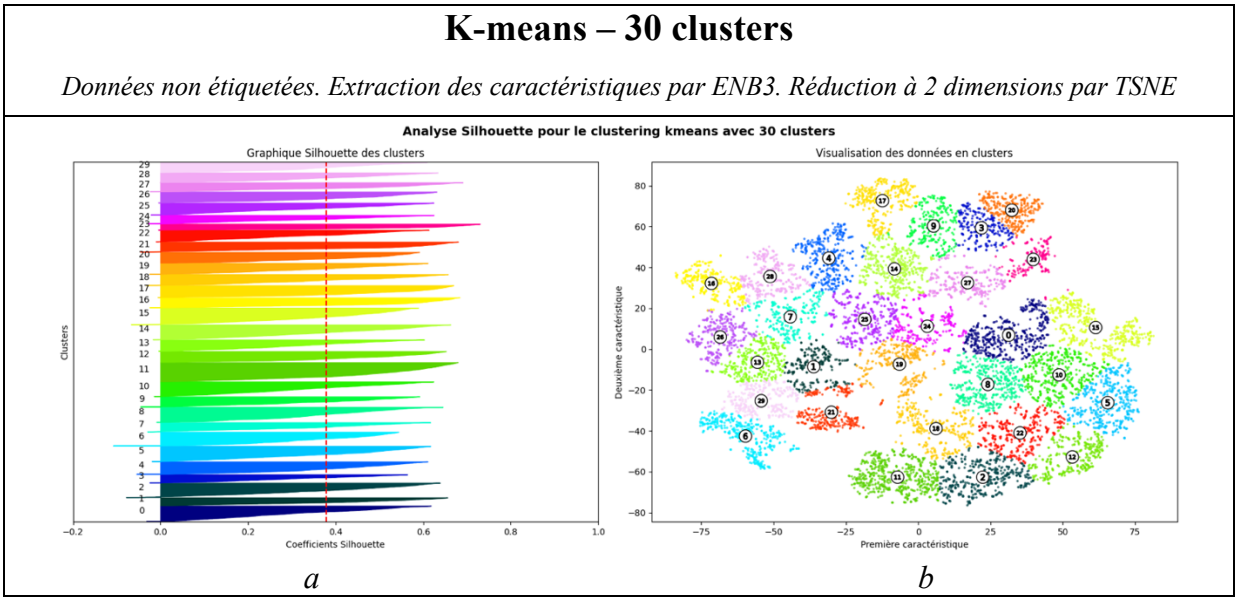


Figure 5.14 Clustering K-means de 30 clusters avec les données non étiquetées: graphique Silhouette (a), graphique 2D en nuages de points (b)

5.3.2.2 Fuzzy C-means

Nous avons poursuivi les expérimentations avec l'algorithme Fuzzy C-means. Tout comme avec l'algorithme K-means, il faut spécifier le nombre de clusters désirés. Nous avons donc débuté en affichant les courbes de la somme de la distance au carré selon le nombre de

clusters. Nous avons réduit le nombre maximum de clusters à 50, car le temps d'exécution de cet algorithme augmente significativement lorsque le nombre de clusters augmente. Les quatre courbes sont présentées à la Figure 5.15.

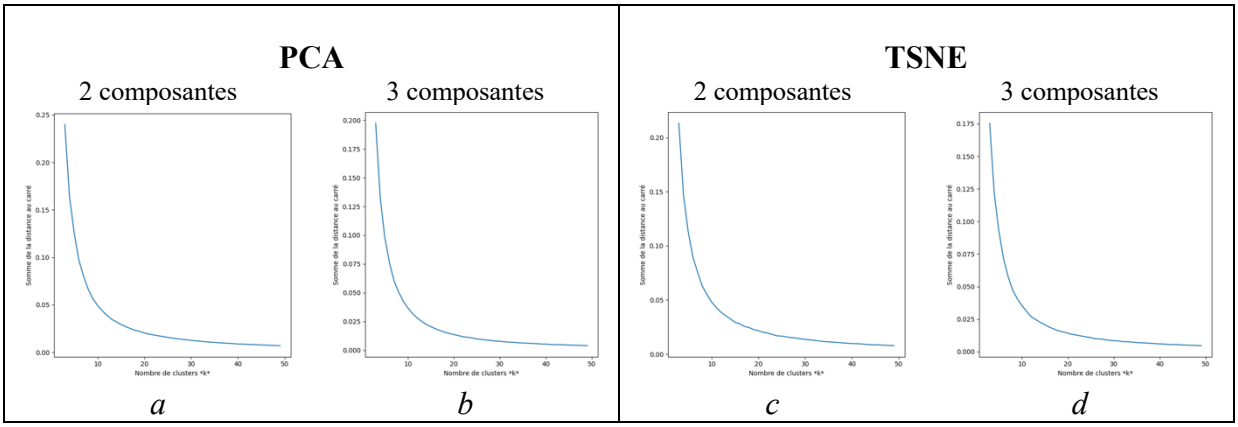


Figure 5.15 Fuzzy C-means : Courbes de la somme de la distance au carré selon le nombre de clusters. PCA avec 2 (a) et 3 composantes (b). TSNE avec 2 (c) et 3 composantes (d)

Nous remarquons que les courbes semblent se stabiliser autour de 30 clusters, ce qui pourrait indiquer que cet algorithme perçoit moins de groupes distincts dans les échantillons. Nous remarquons aussi que les courbes diffèrent peu entre les réductions à deux et à trois dimensions. Nous avons donc appliqué l'algorithme Fuzzy C-means, avec un nombre de 30 clusters, sur ces quatre options pour la réduction de la dimensionnalité. Le graphique de la Figure 5.16 présente les différentes métriques calculées. Tout comme précédemment, nous observons que les métriques sont supérieures lorsque les caractéristiques sont réduites à 2 dimensions par TSNE. Nous avons donc choisi de poursuivre les expérimentations en utilisant cette technique.

Nous avons ensuite évalué l'algorithme Fuzzy C-means avec différentes valeurs pour le nombre de clusters. Les valeurs testées se situent entre 10 et 50, par incrément de 5. Le graphique de la Figure 5.17 présente les différentes métriques calculées.

Fuzzy C-means – 30 clusters

Données étiquetées. Extraction des caractéristiques par ENB3

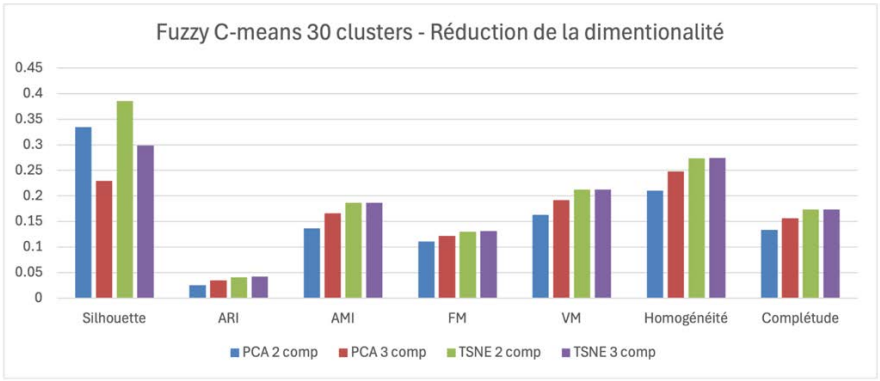


Figure 5.16 Graphique en barres des métriques d'évaluation des clusters Fuzzy C-means, avec différentes méthodes de réduction de la dimensionnalité

Fuzzy C-means

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

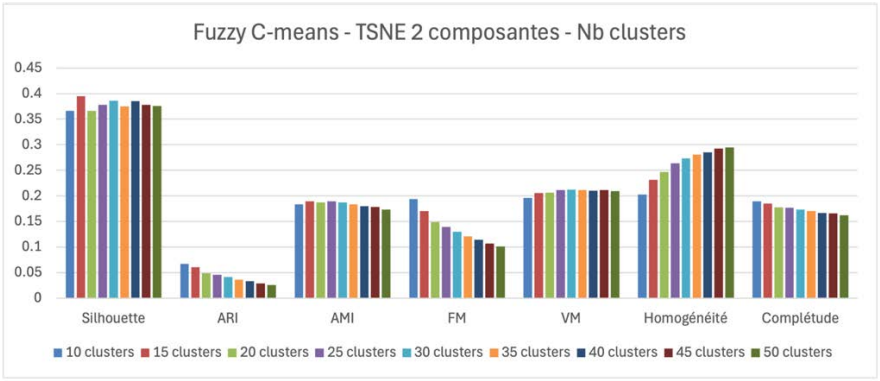


Figure 5.17 Graphique en barres des métriques d'évaluation des clusters Fuzzy C-means avec variation du nombre de clusters

Nous observons un comportement similaire à celui observé lors des expérimentations avec l'algorithme K-means. En se fiant aux coefficients Silhouette moyens, nous constatons qu'un nombre de 15 clusters semble optimal. Nous avons donc généré le graphique Silhouette ainsi

que le graphique en nuages de points pour un clustering de 15 groupes à partir des caractéristiques extraites par ENB3 et réduites à 2 dimensions par TSNE (Figure 5.18).

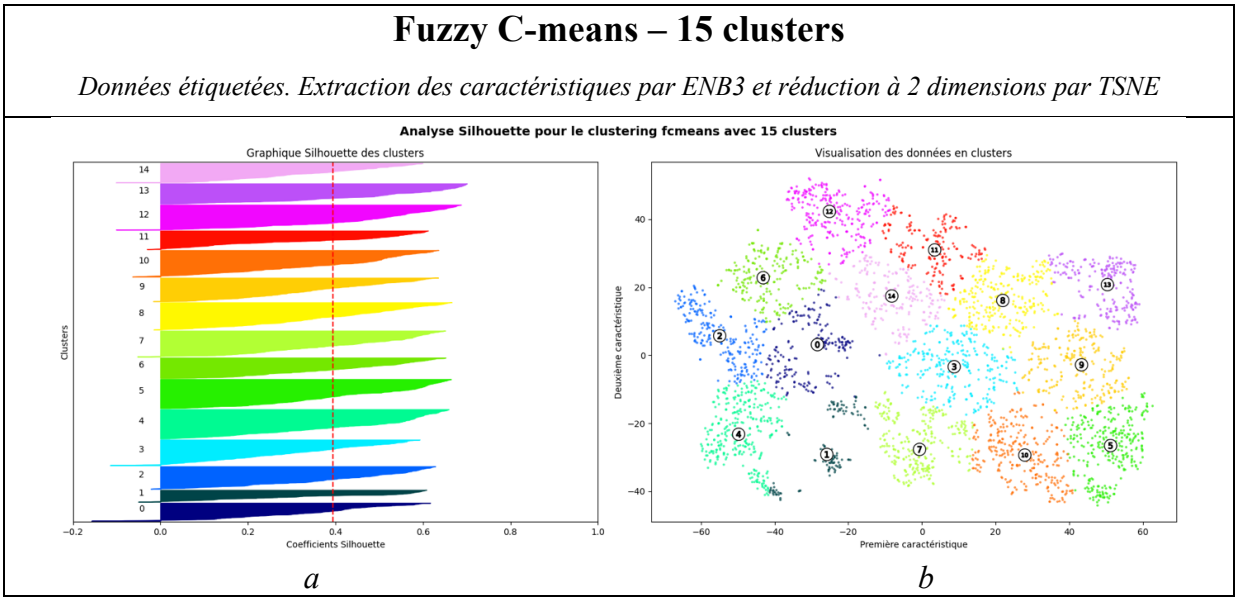


Figure 5.18 Clustering Fuzzy C-means de 15 clusters : graphique Silhouette (a), graphique 2D en nuage de points (b)

Nous avons aussi affiché les images 2D de cinq échantillons appartenant à chacun des clusters en annexe (ANNEXE V, Figure-A V-1). En observant ces images 2D, la qualité des clusters nous semble similaire à celle obtenue par le clustering K-means. Encore une fois, l'avis d'un expert serait utile à cette étape afin d'identifier les échantillons regroupés dans les mauvais clusters.

Puisque le clustering flou calcule une matrice de partition floue indiquant le degré d'appartenance de chaque échantillon à chaque cluster, nous avons effectué quelques expérimentations afin d'évaluer si cette solution pourrait être intéressante. Nous avons donc appliqué un clustering Fuzzy C-means de six groupes, correspondant aux six catégories possibles (excluant la catégorie « inconnu »). L'idée étant d'observer la matrice de partition floue et d'établir un seuil indiquant les clusters valides pour chaque échantillon, établissant ainsi une catégorisation multiétiquette. Un paramètre *m* permet de contrôler le niveau de flou de l'algorithme Fuzzy C-means. Nous avons donc itéré cette expérimentation avec

différentes valeurs pour le paramètre *m*. Différentes métriques d'évaluation des clusters sont présentées dans les graphiques de la Figure 5.19. Pour ces expérimentations, nous avons seulement utilisé les métriques ne nécessitant pas la vérité de terrain, puisque la vérité de terrain correspond aux catégories multiétiquettes et n'est donc pas significative afin d'évaluer ces groupements de six catégories. Deux métriques spécifiques au clustering flou s'ajoutent. Le coefficient de partition (PC, partition coefficient) qui varie de 0 à 1 et qui évalue le niveau de séparation des clusters, avec une valeur de 1 indiquant des clusters clairement séparés. Et le coefficient d'entropie de partition (PEC, partition entropy coefficient), indiquant le niveau de chevauchement (ou de flou) des clusters, avec une valeur plus élevée représentant un chevauchement plus grand. Comme attendu, nous pouvons observer que le coefficient Silhouette moyen et PC diminuent lorsque la valeur de *m* augmente, indiquant que les frontières des clusters sont moins définies; alors que PEC augmente indiquant un plus grand flou.

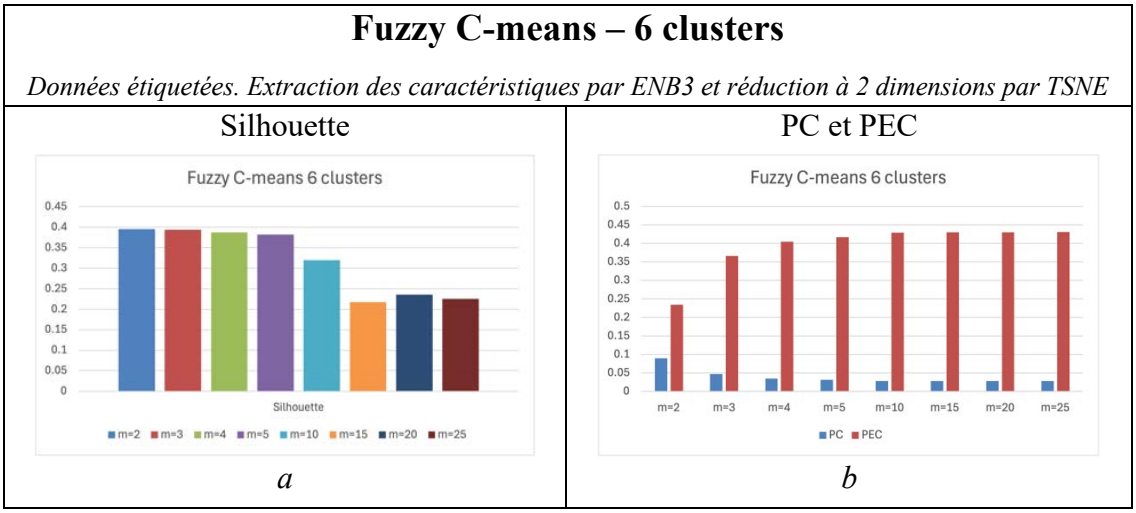


Figure 5.19 Graphiques des métriques d'évaluation des clusters
Fuzzy C-means pour 6 clusters avec différentes valeurs du paramètre *m*.
Graphiques en barres : Silhouette (a), PC et PEC (b)

Le Tableau 5.2 présente une variation des degrés d'appartenance d'un même échantillon à chacun des clusters lorsqu'on modifie le niveau de flou à travers la valeur du paramètre *m*. À titre d'indication, la vérité de terrain de cet échantillon est « parallèle et circulaire ». On remarque que plus la valeur de *m* augmente, plus les différences entre les degrés

d'appartenance sont faibles. Par exemple, lorsque $m = 2$, on voit que l'échantillon appartient à un seul cluster; alors que, lorsque $m = 25$, le même échantillon appartient à pratiquement tous les clusters. Dans la réalité, la majorité des échantillons de nos données devraient posséder entre une et trois étiquettes. Ainsi, nous estimons que $m = 3$ pourrait être préférable, car la différence entre les valeurs d'appartenance pourrait permettre d'établir un seuil facilitant l'atteinte de cette réalité.

Tableau 5.2 Degrés d'appartenance d'un même échantillon, selon la valeur du paramètre *m*

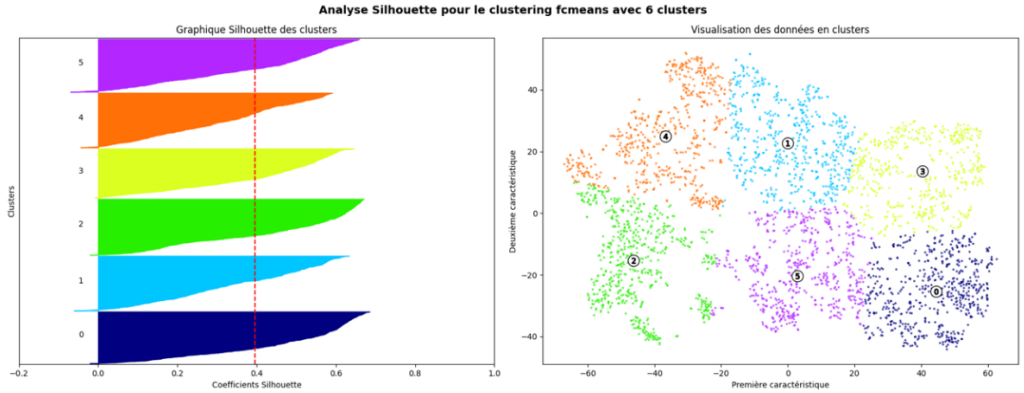
m	Prédiction	C0	C1	C2	C3	C4	C5
2	C2	0,0226	0,0389	0,7853	0,0205	0,0678	0,0649
3	C5	0,1535	0,0821	0,0770	0,1247	0,1069	0,4558
4	C1	0,1372	0,3167	0,1484	0,1092	0,1161	0,1725
5	C3	0,1549	0,1489	0,1328	0,2615	0,1253	0,1765
10	C5	0,1594	0,1533	0,1630	0,1539	0,1754	0,1950
15	C0	0,1822	0,1606	0,1625	0,1648	0,1574	0,1725
20	C2	0,1672	0,1625	0,1782	0,1600	0,1643	0,1678
25	C1	0,1701	0,1748	0,1631	0,1614	0,1644	0,1662

La Figure 5.20 présente les graphiques Silhouette ainsi que les graphiques en nuages de points pour les itérations $m = 2$, $m = 10$ et $m = 25$. On peut observer la forme des clusters s'allonger et se superposer à mesure que la valeur du paramètre *m* augmente, expliquant ainsi la baisse du coefficient Silhouette moyen.

Fuzzy C-means – 6 clusters

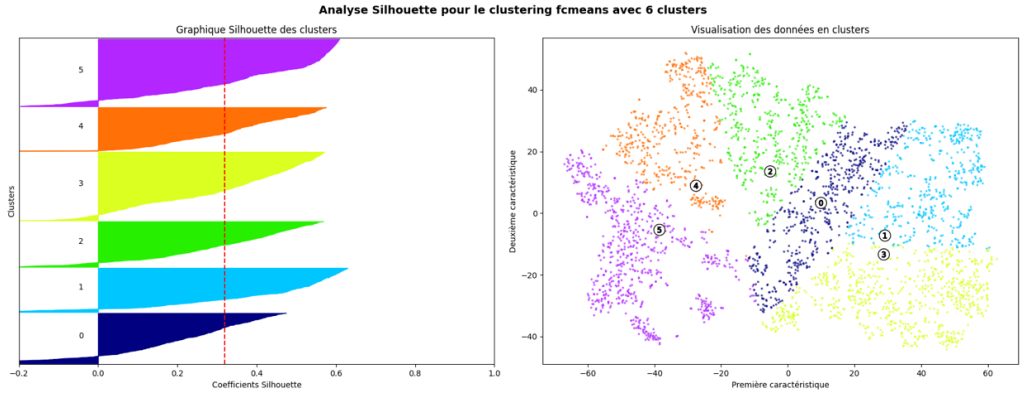
Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

$m = 2$



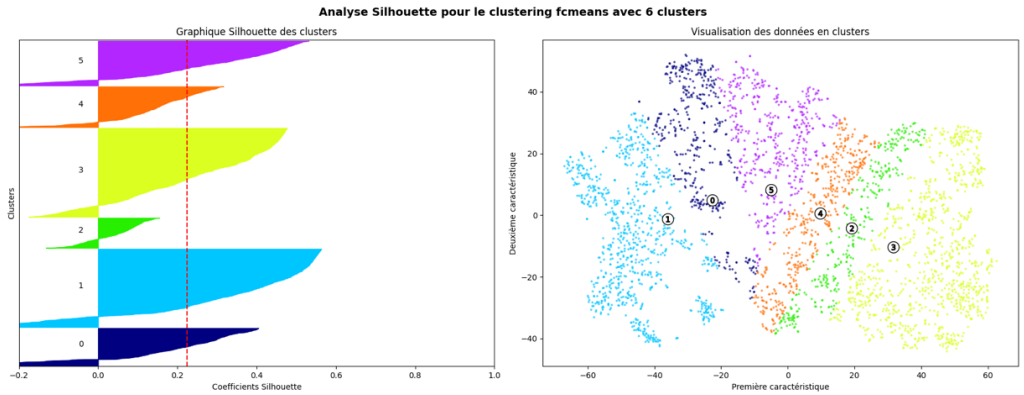
a *b*

$m = 10$



c *d*

$m = 25$



e *f*

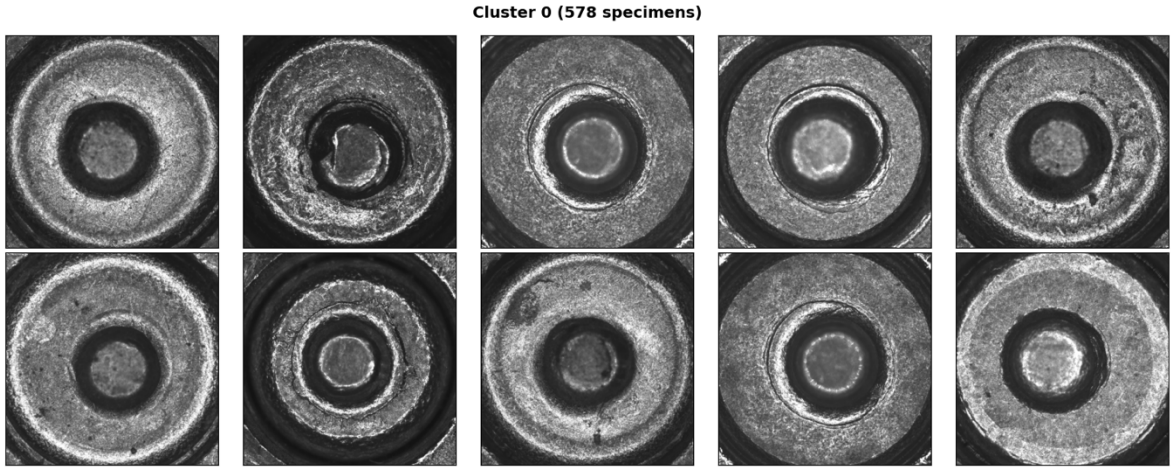
Figure 5.20 Clustering Fuzzy C-means de 6 clusters : graphiques Silhouette (a, c et e), et graphiques 2D en nuages de points (b, d et f)

Afin de poursuivre l'exploration de cette idée, nous avons observé les degrés d'appartenance pour l'algorithme utilisant le paramètre m = 3. Pour cette configuration, nous avons affiché les images 2D de dix échantillons appartenant à chacun des 6 clusters (Figure 5.21). Le Tableau 5.3 présente les degrés d'appartenance correspondant aux échantillons affichés du cluster 1 alors que le Tableau 5.4 présente ceux du cluster 2.

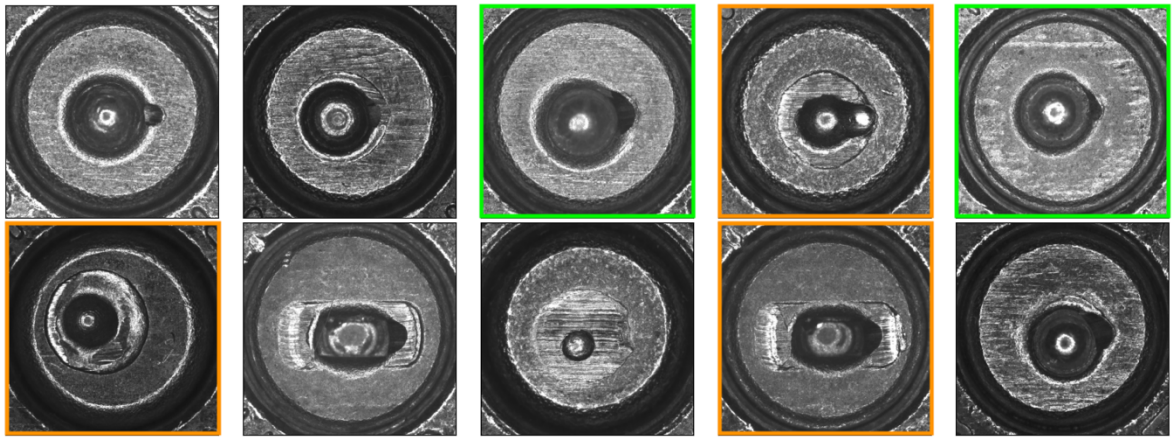
Images 2D des dix premiers échantillons de chaque cluster Fuzzy C-means

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

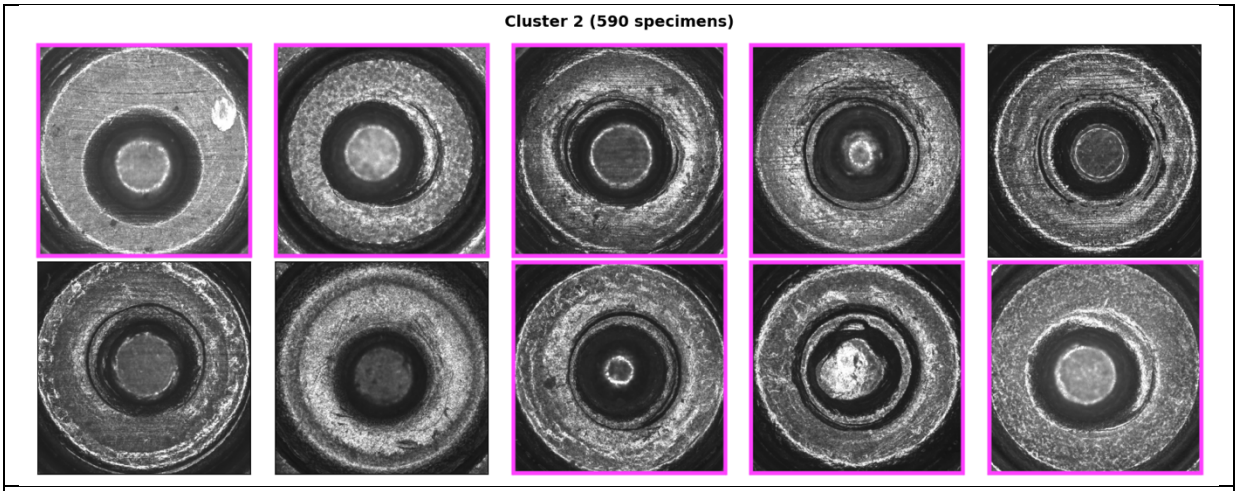
Paramètre : *nb_clusters* = 6, *m* = 3



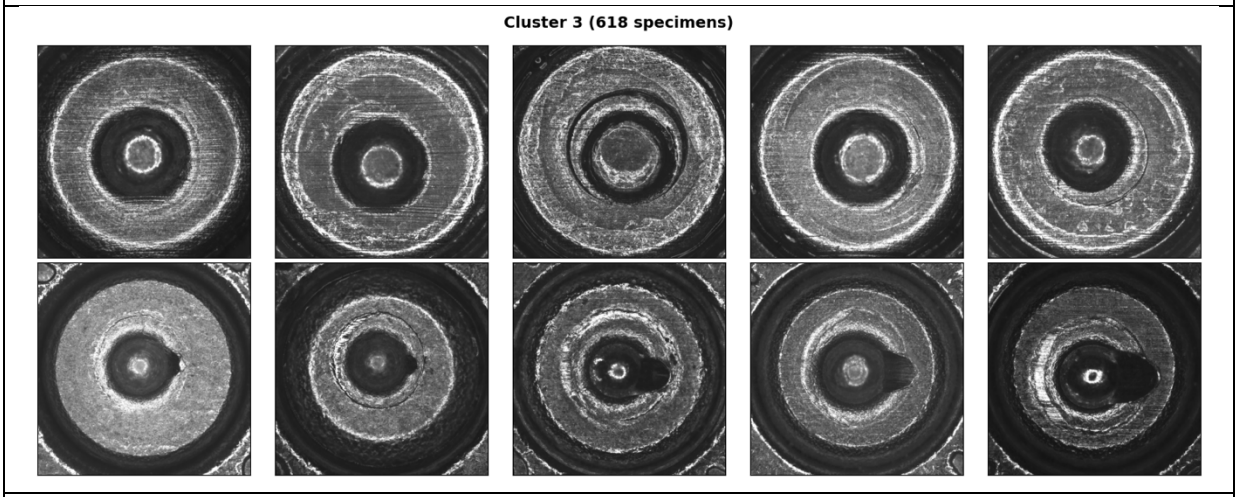
a



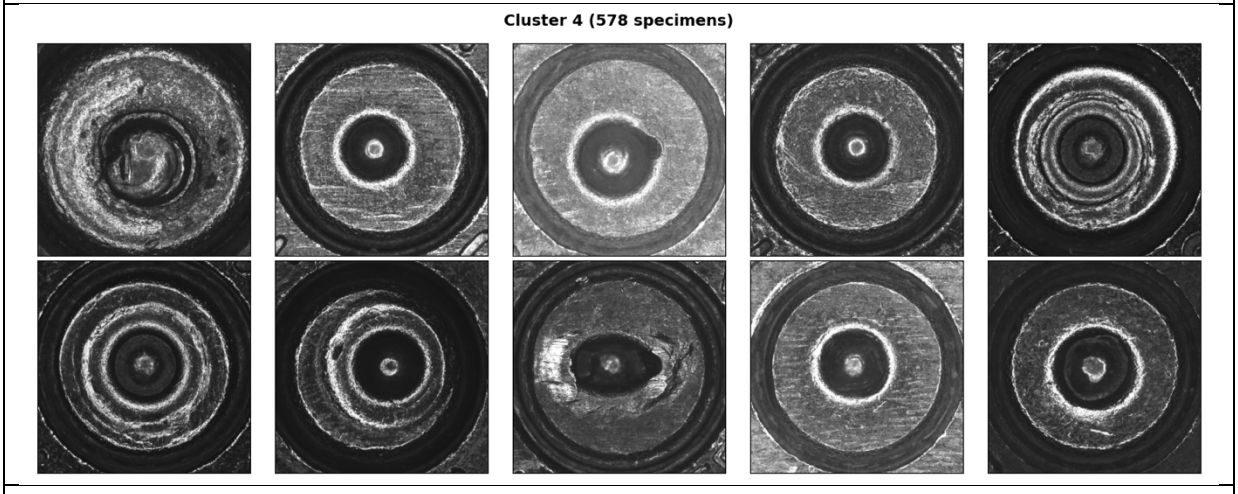
b


Cluster 2 (590 specimens)

c


Cluster 3 (618 specimens)

d


Cluster 4 (578 specimens)

e

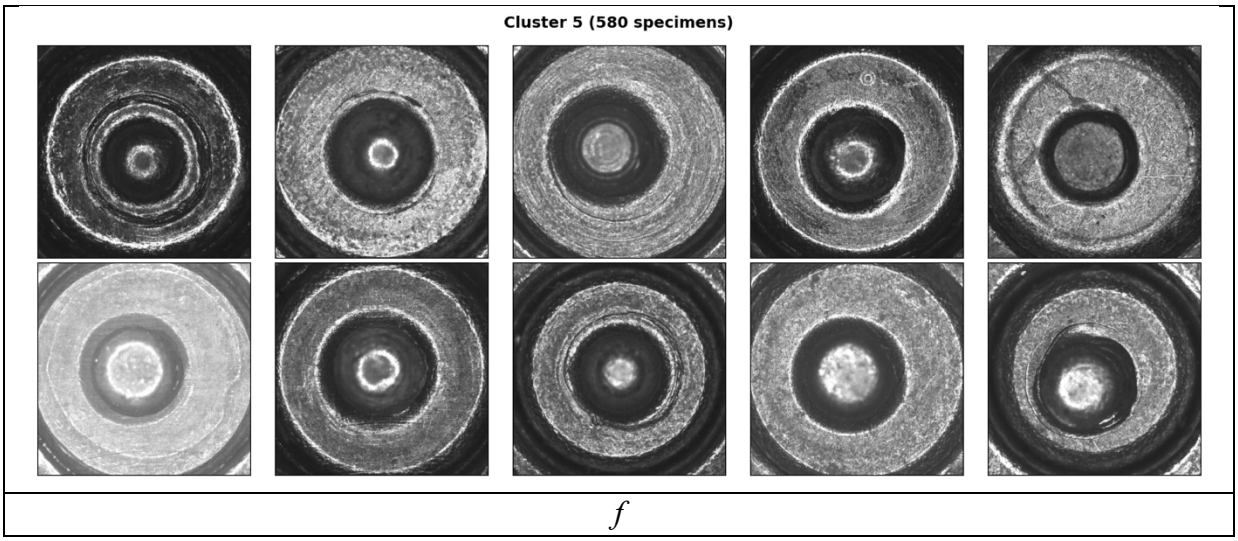


Figure 5.21 Images 2D des dix premiers échantillons de chaque cluster Fuzzy C-means, pour un nombre de clusters de six avec un paramètre m=3

Tableau 5.3 Degrés d'appartenance des dix premiers échantillons du cluster 1

#	C0	C1	C2	C3	C4	C5	GT
1	0,0863	0,3205	0,1368	0,2196	0,1588	0,0780	*hachure, granulaire*
2	0,0530	0,6007	0,1131	0,0952	0,0914	0,0467	*hachure, granulaire*
3	0,0980	0,3299	0,1039	0,1464	0,2447	0,0772	*hachure, granulaire*
4	0,0989	0,3193	0,1993	0,1444	0,1511	0,0869	*inconnu*
5	0,0969	0,3544	0,1035	0,1331	0,2370	0,0751	*inconnu*
6	0,0965	0,3243	0,2023	0,1439	0,1478	0,0852	*parallèle, granulaire*
7	0,0946	0,3771	0,1577	0,1332	0,1576	0,0798	*inconnu*
8	0,0933	0,3736	0,1660	0,1346	0,1528	0,0797	*inconnu*
9	0,0954	0,3447	0,1864	0,1403	0,1501	0,0831	*inconnu*
10	0,0366	0,7300	0,0713	0,0651	0,0652	0,0319	*hachure, granulaire*

Tableau 5.4 Degrés d'appartenance des dix premiers échantillons du cluster 2

#	C0	C1	C2	C3	C4	C5	GT
1	0,0853	0,1460	0,3510	0,2159	0,1046	0,0972	*parallèle, granulaire*
2	0,0897	0,1457	0,2824	0,2680	0,1092	0,1050	*parallèle*
3	0,0865	0,1462	0,3116	0,2495	0,1065	0,0998	*parallèle, granulaire*
4	0,0839	0,1453	0,3585	0,2138	0,1033	0,0952	*parallèle, circulaire*
5	0,0709	0,1396	0,4741	0,1497	0,0901	0,0756	*parallèle, granulaire*
6	0,0809	0,1469	0,4112	0,1725	0,1002	0,0884	*parallèle, granulaire*
7	0,0829	0,1450	0,3766	0,2000	0,1021	0,0933	*granulaire*
8	0,0853	0,1460	0,3512	0,2158	0,1046	0,0971	*granulaire*
9	0,0898	0,1457	0,2819	0,2680	0,1094	0,1053	*granulaire*
10	0,0862	0,1462	0,3149	0,2472	0,1062	0,0993	*parallèle*

Pour le cluster 1, nous pouvons observer que le degré d'appartenance le plus élevé de chaque échantillon correspond bien au cluster 1 (cellules en bleu dans le Tableau 5.3). En observant les images 2D de la Figure 5.21, nous pourrions supposer que ce cluster correspond aux marques de type « hachures ». En plaçant un seuil à 0,1800, nous prédisons que les échantillons appartiennent aussi aux clusters dont le degré d'appartenance est supérieur à 0,1800 (cellules en mauve). Par exemple, les échantillons # 3 et #5 (encadrés et vert dans le cluster 1 de la Figure 5.21) auraient la même catégorie multiétiquette, de même que les échantillons #4, #6 et #9 (encadrés en orange). Cependant, l'assignation d'un seuil est plus difficile, pour le cluster 2 (Tableau 5.4). En assignant un seuil de 0,2100, on trouve que les échantillons #1 à 4 et #8 à 10 (encadrés en rose dans la Figure 5.21) auraient la même catégorie multiétiquette. Ici encore, une analyse plus approfondie incluant la participation d'experts serait probablement nécessaire afin d'évaluer cette technique adéquatement. Toutefois, nous croyons qu'elle pourrait être intéressante afin d'identifier les marques multiétiquettes sur les douilles. Ces expérimentations pourraient faire partie des travaux futurs.

Pour terminer avec les expérimentations de l'algorithme Fuzzy C-means, nous l'avons appliqué sur les **données non étiquetées**. Nous avons utilisé un nombre de six clusters et une valeur de paramètre m = 3. La Figure 5.22 présente le graphique Silhouette ainsi que le graphique en nuages de points pour ce clustering à partir des caractéristiques extraites par ENB3 et réduites à 2 dimensions par TSNE.

Fuzzy C-means – 6 clusters

Données non étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

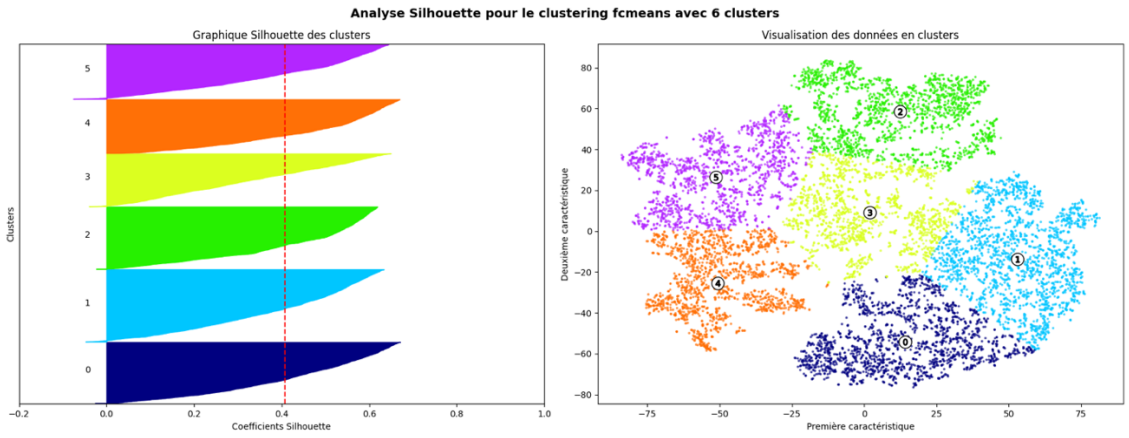


Figure 5.22 Clustering Fuzzy C-means de 6 clusters avec les données non étiquetées: graphique Silhouette (a), graphique 2D en nuage de points (b)

5.3.2.3 DBSCAN

Avec l'algorithme DBSCAN, deux paramètres sont particulièrement importants pour la formation des clusters. Le paramètre *eps*, qui représente la distance maximale entre deux points afin qu'ils puissent être considérés comme étant le dans le voisinage l'un de l'autre; et le *min_samples* qui indique le nombre minimum d'échantillons devant appartenir au même voisinage afin qu'un point puisse être déclaré comme étant un point noyau (core). En pratique, plusieurs itérations de l'algorithme sont nécessaires afin d'identifier la combinaison optimale pour les valeurs de ces paramètres, et cette combinaison peut varier lorsque les données sont modifiées (par exemple avec un changement de la méthode d'extraction ou de réduction). Pour cette raison, il est difficile de comparer les différentes méthodes d'extraction

des caractéristiques et de réduction de la dimensionnalité, comme nous l'avons fait précédemment en utilisant des paramètres fixes. Pour cette section, nous présentons les résultats des expérimentations effectuées sur les caractéristiques extraites par le modèle ENB3 avec une réduction à 2 dimensions par TSNE. L'algorithme a été appliqué en utilisant la métrique de distance Euclidienne.

DBSCAN

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

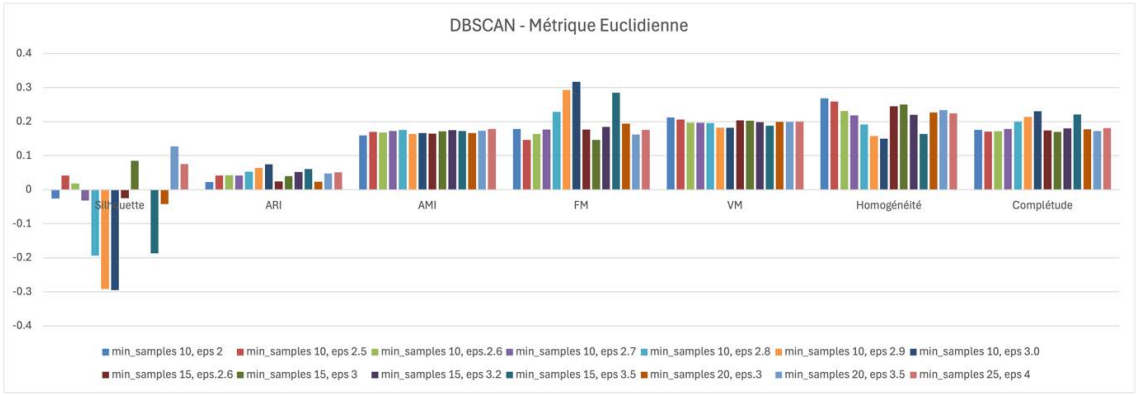


Figure 5.23 Graphique en barres des métriques d'évaluation des clusters DBSCAN, avec différentes combinaisons de paramètres

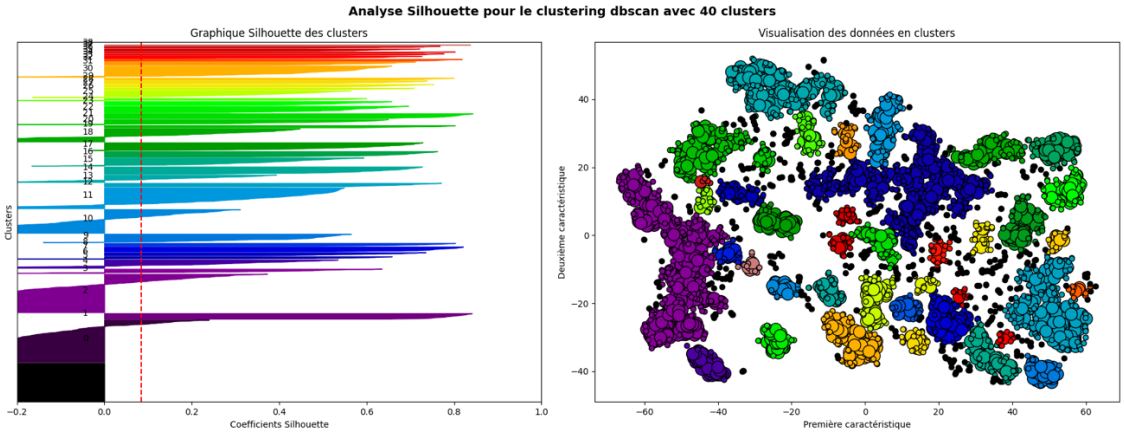
Le graphique de la Figure 5.23 présente les métriques calculées pour différentes combinaisons des paramètres *min_samples* et *eps*. Nous remarquons que les valeurs des coefficients Silhouette sont moins élevées que celles pour les clusters formés avec les algorithmes précédents. Cette baisse s'explique par les variations des tailles et des formes des clusters; ainsi que par la présence des valeurs aberrantes, qui sont éparpillées un peu partout. Pour ces raisons, cette métrique n'est pas particulièrement adaptée pour l'évaluation d'un clustering basé sur la densité. Nous l'avons tout de même utilisée pour nous guider dans la comparaison des résultats et le choix des paramètres. Ainsi, nous observons que trois configurations de paramètres se distinguent du lot pour la valeur du coefficient Silhouette moyen : *min*_samples = 15 avec *eps* = 3, *min_samples* = 20 avec *eps* = 3,5, et *min_samples* = 25 avec *eps* = 4.

La Figure 5.24 (a à f) présente les graphiques Silhouette ainsi que les graphiques en nuages de points pour ces trois expérimentations. Les points correspondants aux valeurs aberrantes sont regroupés dans un même cluster, identifié par -1. Dans les graphiques, ces points sont représentés en noir. Nous remarquons les formes et les dimensions variées des clusters, qui pourraient mieux correspondre à la répartition de nos données dans les différentes catégories.

DBSCAN

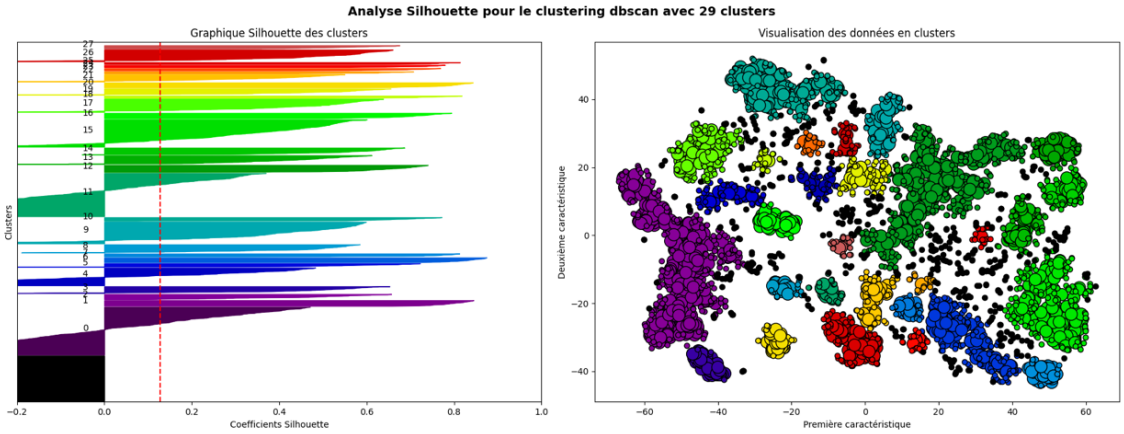
Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

Paramètres : *min_samples = 15, eps = 3*
Résultat : 40 clusters



a *b*

Paramètres : *min_samples = 20, eps = 3,5*
Résultat : 29 clusters



c *d*

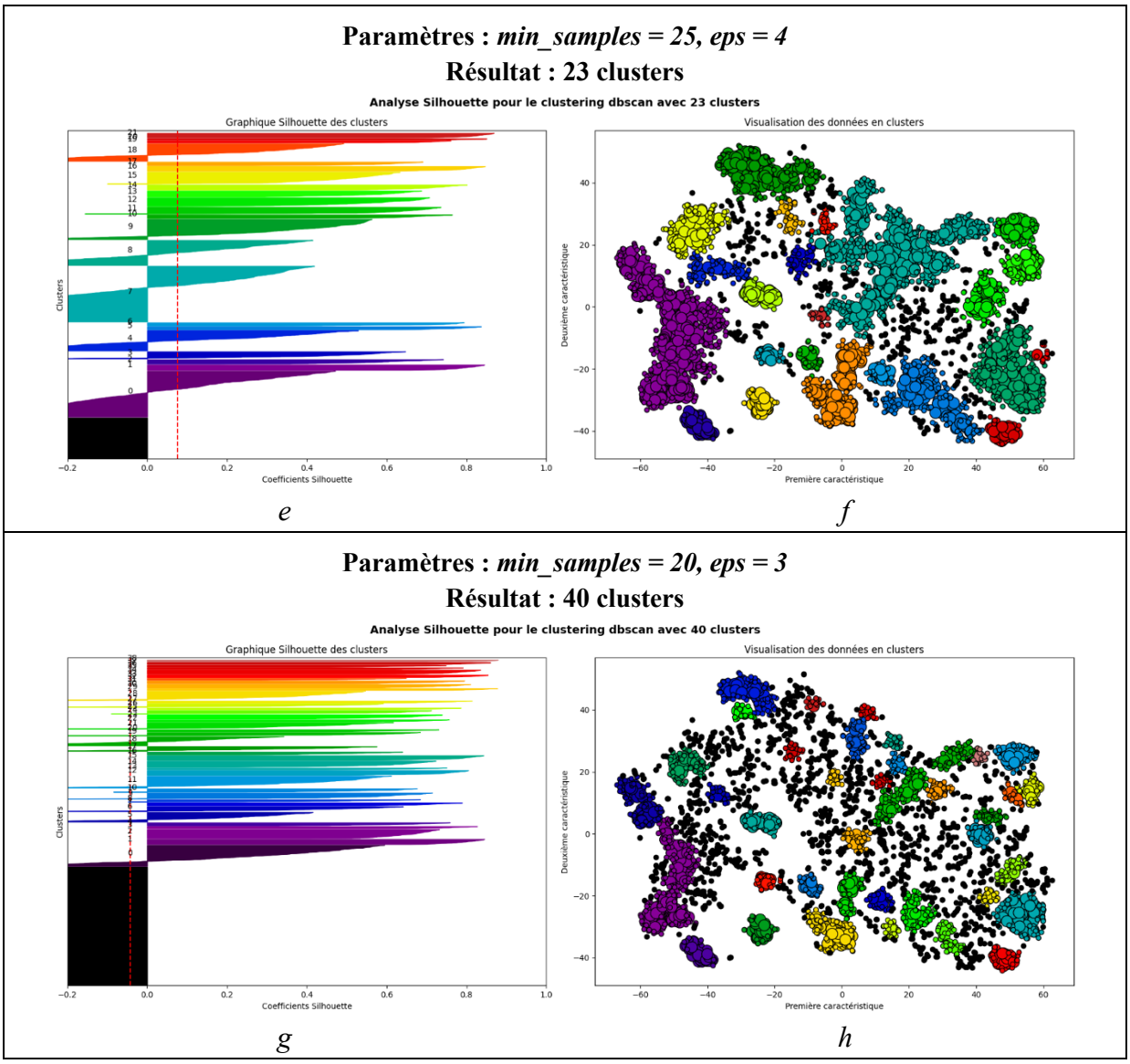


Figure 5.24 Clustering DBSCAN pour différentes combinaisons de paramètres : graphiques Silhouette (a, c, e et g) et graphiques 2D en nuage de points (b, d, f et h)

Nous remarquons aussi que plus la valeur des paramètres augmente, plus les groupes sont larges et le nombre de clusters diminue. À titre de comparaison, nous avons ajouté les graphiques pour une configuration de paramètres dont les métriques présentent une performance inférieure : *min_samples* = 20 avec *eps* = 3. Pour cette configuration, la valeur du coefficient Silhouette est négative (Figure 5.24 g, h). Nous remarquons que lorsque les

paramètres sont moins bien adaptés, la taille des clusters diminue et la quantité de valeurs aberrantes augmente.

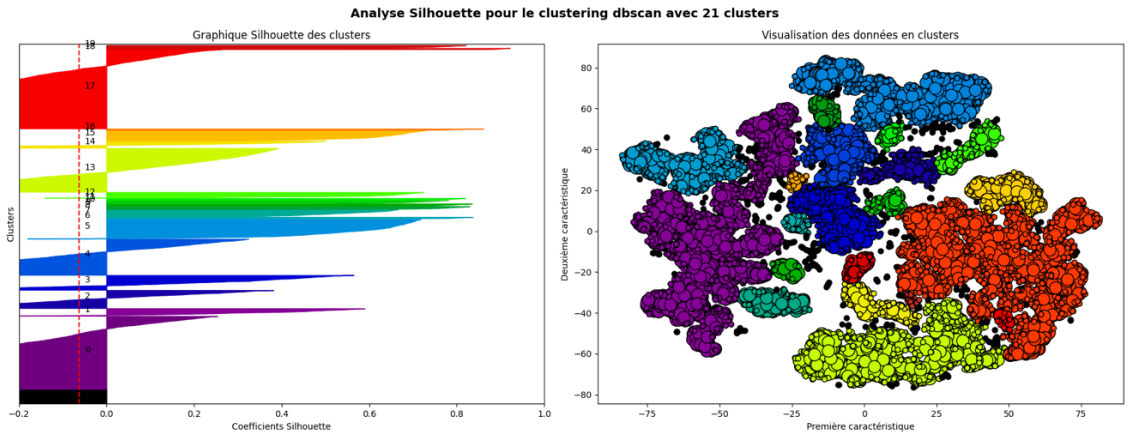
En observant les images 2D des échantillons de chaque cluster, nous constatons que la répartition des groupes semble plus proche de celle de nos catégories de marques qu'avec les expérimentations précédentes. Cependant, nous observons la présence de marques sur les images catégorisées comme étant des valeurs aberrantes, et nous croyons qu'elles auraient pu appartenir à des clusters plus pertinents. Les images 2D des cinq premiers échantillons appartenant à chacun des 29 clusters de la configuration avec les paramètres *min_samples* = 20 avec *eps* = 3,5 ainsi que les images 2D des 15 premiers échantillons identifiés comme étant des valeurs aberrantes pour cette même configuration sont présentés en annexe (ANNEXE VI, Figure-A VI-1 et Figure-A VI-2).

DBSCAN

Données non étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

Paramètres : *min_samples = 25, eps = 3,5*

Résultat : 21 clusters



a *b*

Figure 5.25 Clustering DBSCAN avec les données non étiquetées : graphique Silhouette (a) et graphique 2D en nuage de points (b)

Finalement, nous avons appliqué l'algorithme DBSCAN sur les **données non étiquetées**. Nous avons utilisé les paramètres *min_samples* = 25 avec *eps* = 3,5, avec lesquels nous avons

obtenu de bons résultats sur les données étiquetées. La Figure 5.25 présente le graphique Silhouette ainsi que le graphique en nuages de points pour cette expérimentation faite à partir des caractéristiques extraites par ENB3 et réduites à 2 dimensions par TSNE. Nous observons que le nombre de clusters diminue, ainsi que la métrique Silhouette. Nous remarquons aussi plusieurs groupes qui se chevauchent.

5.3.2.4 HDBSCAN

HDBSCAN est un algorithme hiérarchique qui applique l'algorithme DBSCAN plusieurs fois en variant la valeur du paramètre *eps* pour trouver la meilleure solution possible. Ce paramètre n'est donc pas à configurer. Nous avons varié les paramètres *min_cluster_size*, qui indique le nombre d'échantillons minimum dans un groupe pour pouvoir être considéré comme étant un cluster et *min_samples*, qui indique le nombre de voisins à considérer dans le calcul des distances. Dans cette section, nous présentons les résultats des expérimentations effectuées avec les caractéristiques extraites par le modèle ENB3 et réduites à 2 dimensions par TSNE. L'algorithme a été appliqué en utilisant les métriques de distance Euclidienne et cosine.

HDBSCAN – Distance Euclidienne

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

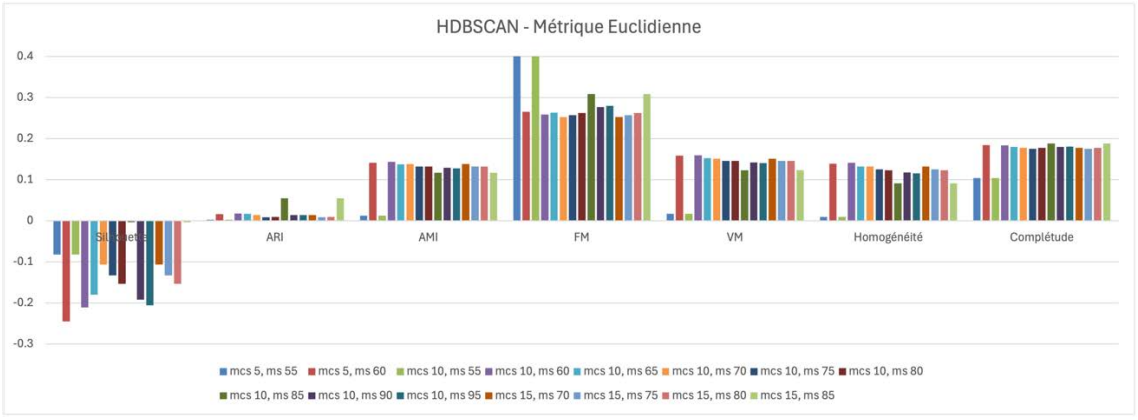


Figure 5.26 Graphique en barres des métriques d'évaluation des clusters HDBSCAN utilisant la distance Euclidienne, avec différentes combinaisons de paramètres

Le graphique de la Figure 5.26 présente les métriques calculées pour différentes combinaisons des paramètres *min_cluster_size et min_samples*, lorsque la métrique de distance Euclidienne est utilisée. Nous observons qu'avec cette métrique de distance, le coefficient Silhouette moyen est souvent inférieur à zéro. Les autres métriques varient et nous observons quelques configurations qui se distinguent avec la métrique FM. Autrement, nous n'observons pas de configuration qui se distingue particulièrement.

Nous trouvons un bon compromis avec les deux combinaisons suivantes : *min_cluster_size* 10, *min_samples* 85 et *min_cluster_size* 15, *min_samples* 85, qui affichent les coefficients Silhouette moyens les moins bas. Cependant, en observant les graphiques Silhouette et les graphiques en nuages de point de ces deux configurations, nous observons que les clusters formés semblent moins intéressants. La Figure 5.27 (a, b) montre ces graphiques pour la configuration *min_cluster_size* 10, *min_samples* 85. La Figure 5.27 (c, d) montre ces graphiques pour la configuration *min_cluster_size* 5, *min_samples* 57, qui nous semblent plus intéressants, puisqu'il affiche un plus grand nombre de groupes. Malgré tout, nous ne croyons pas que ce soit une solution optimale vu la quantité de valeurs aberrantes observées (2 160).

Le graphique de la Figure 5.28 présente les métriques calculées pour différentes combinaisons des paramètres *min_cluster_size et min_samples*, lorsque la métrique de distance cosine est utilisée.

Pour ces itérations de l'algorithme, nous observons que les coefficients Silhouette moyens ont tendance à avoir des valeurs plus élevées. Ce qui nous indique que les clusters sont mieux formés et se distinguent davantage les uns des autres. Avec cette métrique, nous remarquons que les combinaisons *min_cluster_size* 10, *min_samples* 75 et *min_cluster_size* 15, *min_samples* 75 se distinguent avec les valeurs de coefficient Silhouette moyen les plus élevées.

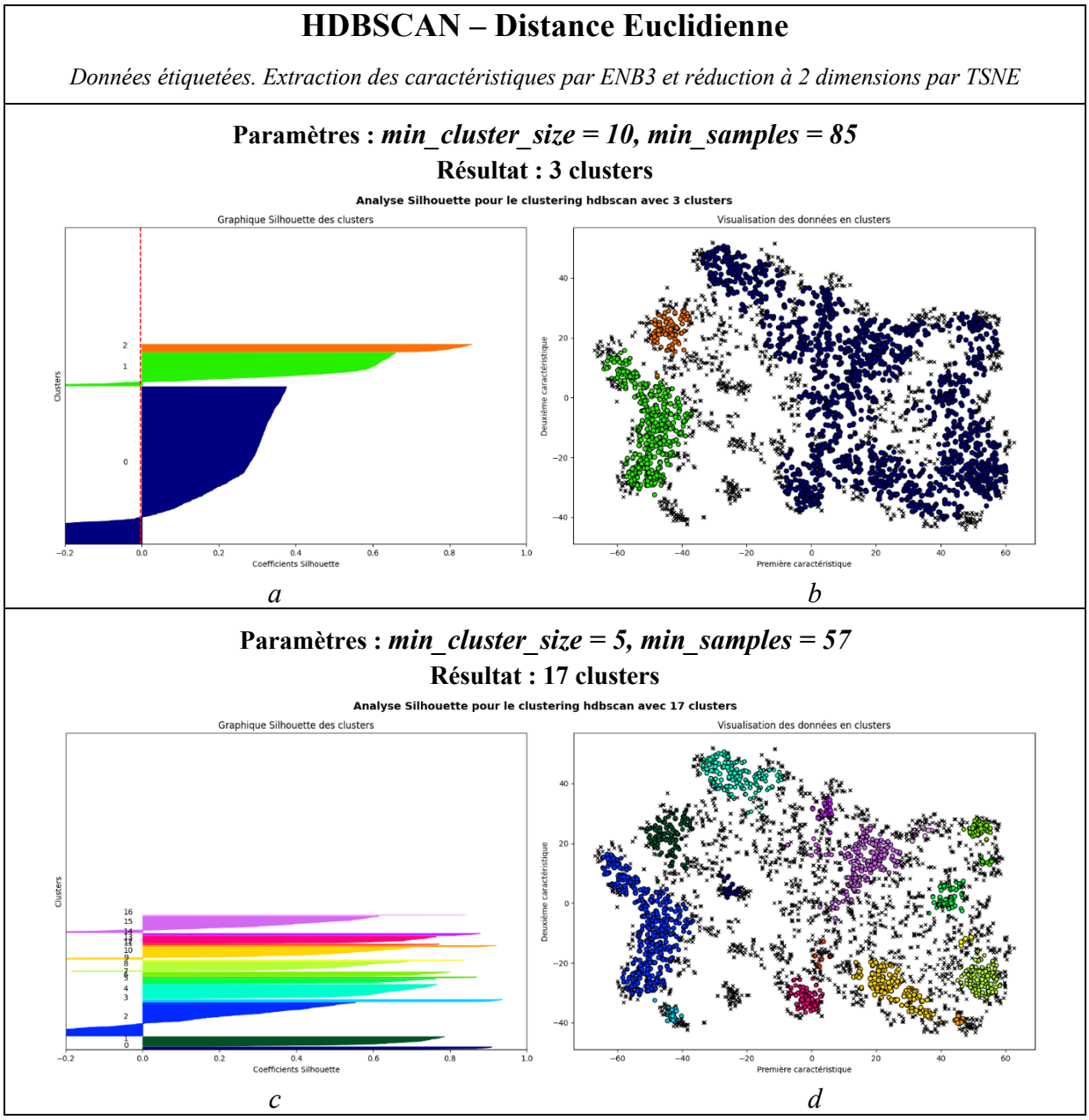


Figure 5.27 Clustering HDBSCAN utilisant la métrique de distance Euclidienne, pour différentes combinaisons de paramètres : graphiques Silhouette (a, c) et graphiques 2D en nuage de points (b, d)

HDBSCAN – Distance Cosine

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

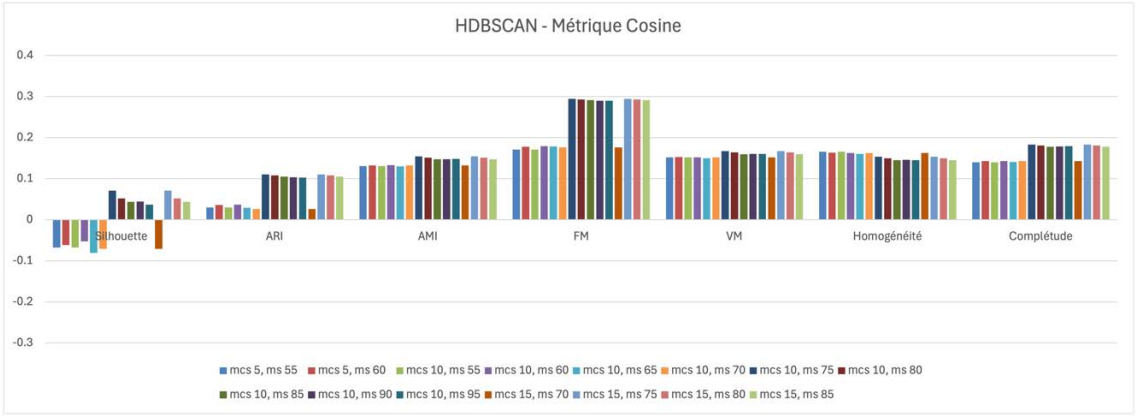


Figure 5.28 Graphique en barres des métriques d'évaluation des clusters HDBSCAN utilisant la distance cosine, avec différentes combinaisons de paramètres

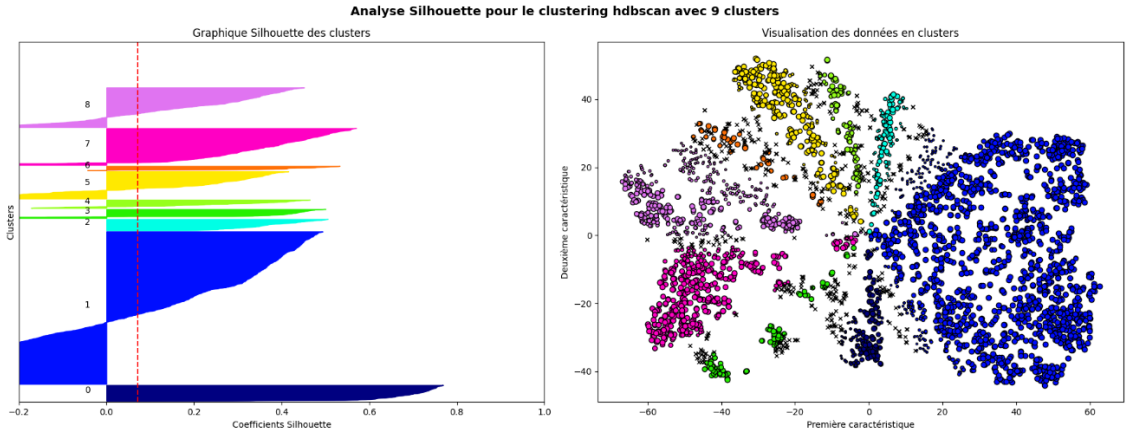
En observant les graphiques de différentes configurations, nous remarquons que cette métrique favorise les groupes larges. En observant les graphiques Silhouette et les graphiques en nuage de points de toutes les expérimentations, nous remarquons que les clusters formés avec la configuration de paramètres *min_cluster_size* 5, *min_samples* 55 semble plus intéressante. La Figure 5.29 (a, b) montre ces graphiques pour la configuration *min_cluster_size* 10, *min_samples* 75. La Figure 5.29 (c, d) montre ces graphiques pour la configuration *min_cluster_size* 5, *min_samples* 55, qui malheureusement possède toujours un nombre de valeurs aberrantes important (1 026). En observant les images 2D des clusters de cette dernière configuration de paramètres, nous avons tendance à croire que les échantillons sont mieux répartis dans les clusters selon les types de marques, mais nous trouvons l'évaluation extrêmement difficile sans l'avis d'un expert. Les images 2D des cinq premiers échantillons appartenant à chacun des 19 clusters de cette dernière configuration, ainsi que les images 2D des 15 premiers échantillons identifiés comme étant des valeurs aberrantes sont présentées en annexe (ANNEXE VII, Figure-A VII-1 et Figure-A VII-2).

HDBSCAN – Distance Cosine

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

Paramètres : *min_cluster_size = 10, min_samples = 75*

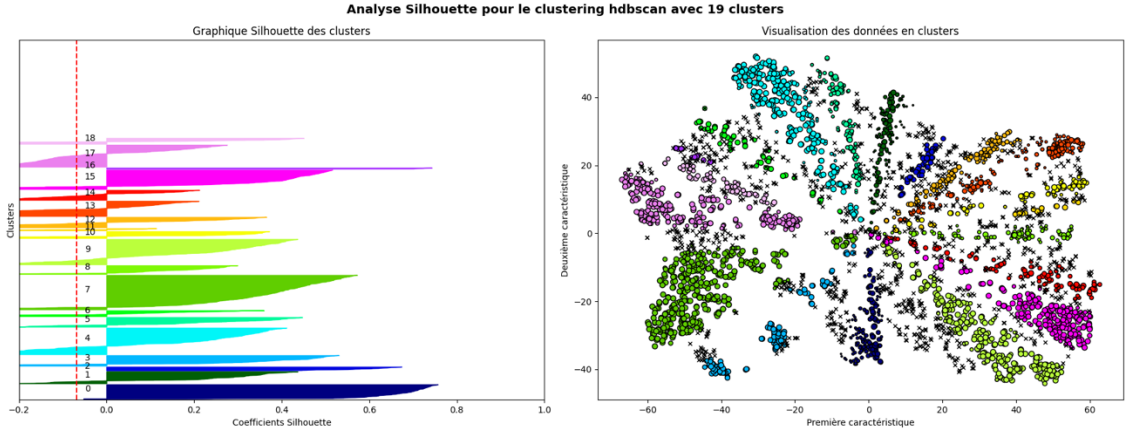
Résultat : 9 clusters



a *b*

Paramètres : *min_cluster_size = 5, min_samples = 55*

Résultat : 19 clusters



c *d*

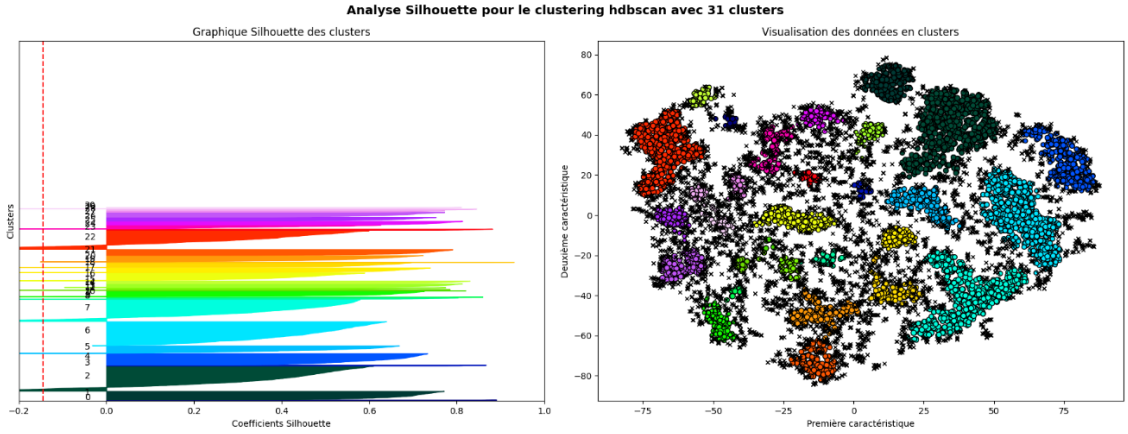
Figure 5.29 Clustering HDBSCAN utilisant la métrique de distance cosine, pour différentes combinaisons de paramètres : graphiques Silhouette (a, c) et graphiques 2D en nuage de points (b, d)

HDBSCAN

Données non étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

Paramètres : *metric = Euclidean, min_cluster_size = 5, min_samples = 55*

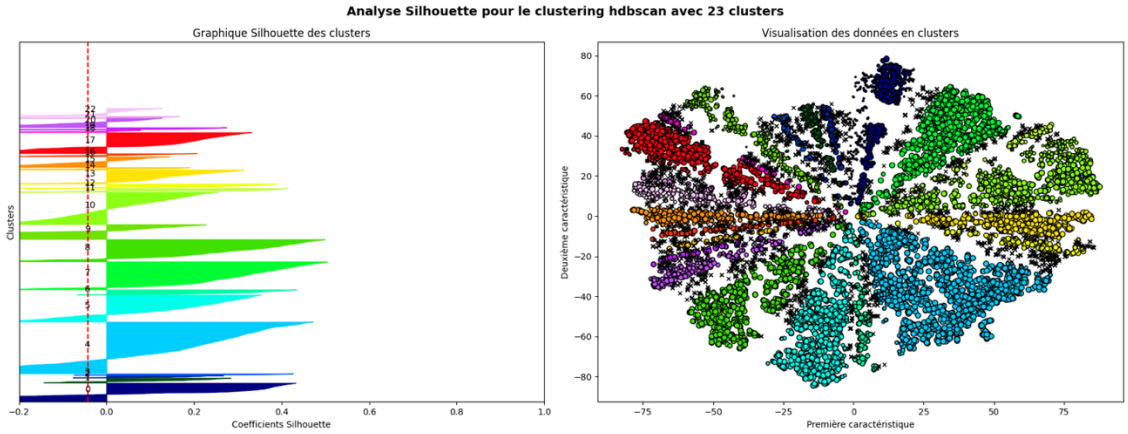
Résultat : 31 clusters



a *b*

Paramètres : *metric = cosine, min_cluster_size = 10, min_samples = 95*

Résultat : 23 clusters



c *d*

Figure 5.30 Clustering HDBSCAN avec les données non étiquetées : graphiques Silhouette (a, c) et graphiques 2D en nuage de points (b, d)

Finalement, nous avons appliqué l'algorithme HDBSCAN sur les **données non étiquetées**. Nous avons testé plusieurs configurations de paramètres, avec les deux métriques de distance. Nous remarquons qu'un grand nombre d'échantillons sont considérés comme étant des valeurs aberrantes, particulièrement avec la métrique de distance Euclidienne. D'un autre côté, nous avons l'impression que les images 2D des échantillons dans chaque cluster

semblent moins cohérentes avec la métrique cosine. La Figure 5.30 présente les graphiques Silhouette ainsi que les graphiques en nuages de points pour deux de ces expérimentations, exécutées à partir des caractéristiques extraites par ENB3 et réduites à 2 dimensions par TSNE.

5.3.2.5 OPTICS

Les principaux paramètres à optimiser pour l'algorithme OPTICS sont *min_cluster_size*, qui indique le nombre d'échantillons minimum dans un groupe pour pouvoir être considéré comme étant un cluster et *min_samples*, qui indique le nombre de voisins minimum qu'un point doit avoir pour être considéré comme un point noyau (core). De plus, un paramètre *cluster_method* permet de sélectionner la méthode de clustering. Deux options sont possibles: la méthode xi avec un paramètre *xi* pour configurer la pente; et la méthode dbscan avec un paramètre *eps* pour configurer la distance maximale entre deux points, afin qu'ils puissent être considérés comme étant le dans le voisinage l'un de l'autre. La métrique de distance est aussi configurable, permettant ainsi une variété d'options. Dans cette section, nous présentons les résultats des expérimentations effectuées à partir des caractéristiques extraites par le modèle ENB3 et réduites à 2 dimensions par TSNE.

Pour chaque méthode de clustering et métrique de distance, les paramètres *min_cluster_size* et *min*_samples, ainsi que le paramètre spécifique à la méthode de clustering (*xi* ou *eps*), doivent être optimisés. Puisque le nombre de configurations à tester est élevé, nous avons d'abord effectué des observations préliminaires en affichant seulement les graphiques en nuages de points des clusters de différentes configurations de paramètres. À cette étape, nous recherchions des graphiques illustrant un nombre de 15 à 30 clusters, assez rapprochés les uns des autres afin d'inclure un maximum d'échantillons et ainsi minimiser la quantité de valeurs aberrantes. D'après ces itérations préliminaires, dont les résultats ne sont pas présentés dans cette thèse, nous avons choisi d'utiliser les paramètres présentés dans le Tableau 5.5.

Tableau 5.5 Paramètres pour l'algorithme OPTICS

Paramètre	**Valeurs**
cluster_method	dbscan
eps	3,1; 3,2
metric	Euclidienne, Minkowski
min_cluster_size (mcs)	5
min_samples (ms)	13, 14, 15

Le graphique de la Figure 5.31 présente les métriques calculées pour les itérations avec la métrique Euclidienne. Tout comme avec les autres algorithmes basés sur la densité évalués précédemment (DBSCAN, HDBSCAN), la métrique Silhouette est difficile à interpréter à cause de la présence des valeurs aberrantes et des formes irrégulières des clusters. Nous remarquons que la valeur du coefficient Silhouette moyen est plus élevée avec les configurations *eps* = 3,1, *min_samples* = 14 et *eps* = 3,1, *min_samples* = 15; alors que la valeur de la métrique FM est plus élevée pour la configuration *eps* = 3,2, *min_samples* = 13.

OPTICS – Métrique Euclidienne

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

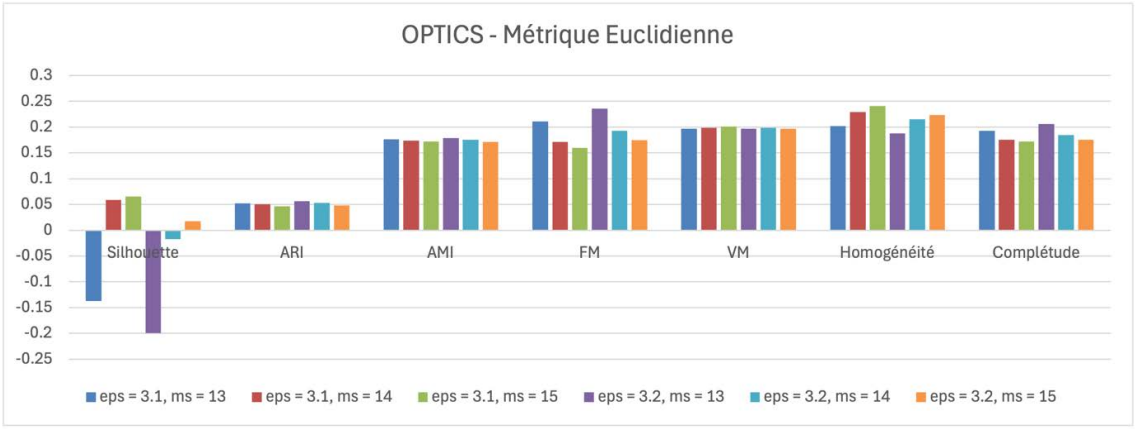


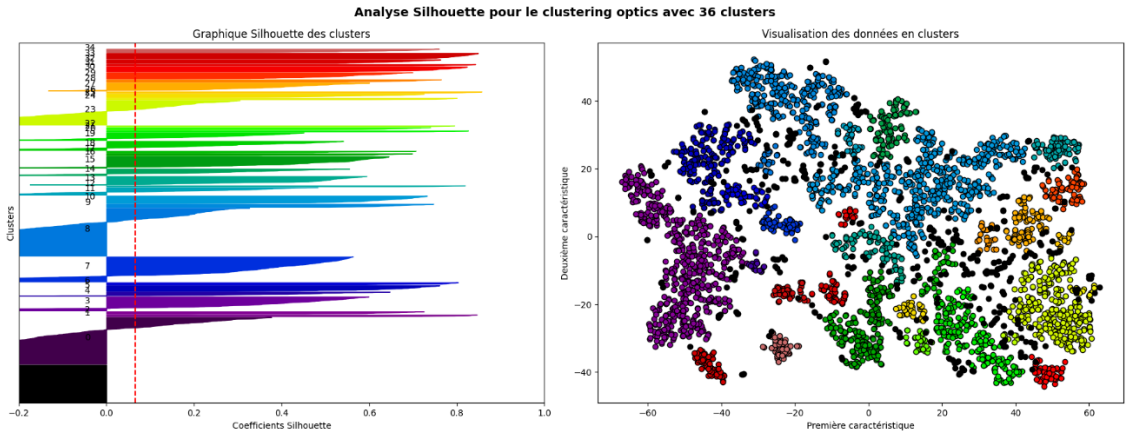
Figure 5.31 Graphique en barres des métriques d'évaluation des clusters OPTICS utilisant la métrique de distance Euclidienne, avec différents paramètres

OPTICS – Métrique Euclidienne

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

Paramètres : *eps = 3,1, min_cluster_size = 5, min_samples = 15*

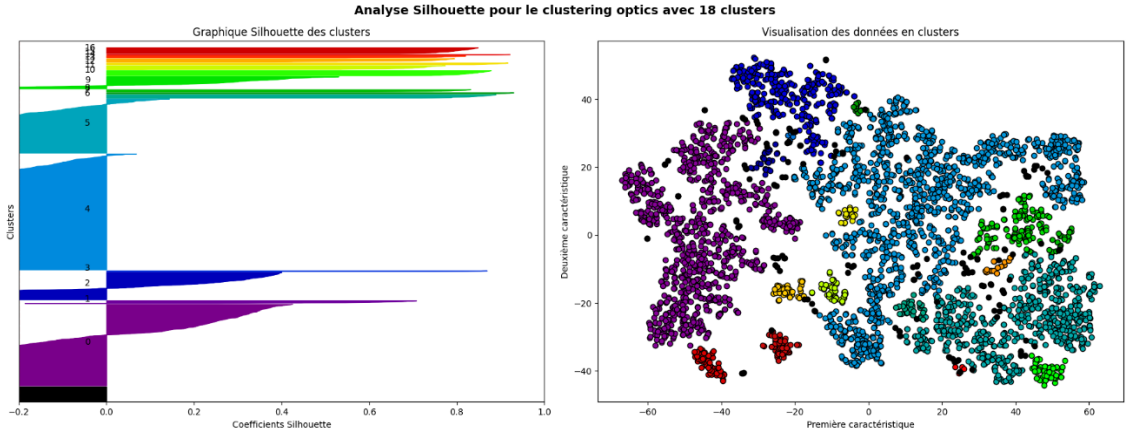
Résultat : 36 clusters



a *b*

Paramètres : *eps = 3,2, min_cluster_size = 5, min_samples = 13*

Résultat : 18 clusters



c *d*

Figure 5.32 Clustering OPTICS utilisant la métrique de distance Euclidienne : graphiques Silhouette (a, c) et graphiques 2D en nuage de points (b, d)

En observant les graphiques Silhouette et les graphiques en nuages de points de chacune des configurations, nous observons que les résultats produits par la combinaison *eps* = 3,2, *min_samples* = 13 semblent plus intéressants, principalement à cause de la présence de trois groupes larges, qui pourraient correspondre aux catégories plus communes, ainsi qu'une faible quantité de valeurs aberrantes. Les graphiques de deux de ces trois configurations

(*eps* = 3,1, *min_samples* = 15 et *eps* = 3,2, *min_samples* = 13) sont présentés dans la Figure 5.32. On peut y observer qu'en augmentant la valeur du paramètre *eps,* la taille de certains clusters augmente. En observant les images 2D des échantillons dans les différents clusters, nous n'observons pas d'amélioration majeure dans la répartition des échantillons parmi les clusters.

La Figure 5.33 présente les métriques calculées pour les itérations avec la métrique de distance Minkowski. Les valeurs sont très proches de celles obtenues avec la métrique de distance Euclidienne. Les mêmes observations s'appliquent donc. En examinant les graphiques Silhouette et les graphiques en nuages de points de chacune des configurations, nous remarquons que les résultats produits par la combinaison *eps* = 3,2, *min_samples* = 14 semble la plus intéressante. Les graphiques de cette configuration sont présentés dans la Figure 5.34.

OPTICS – Métrique Minkowski

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

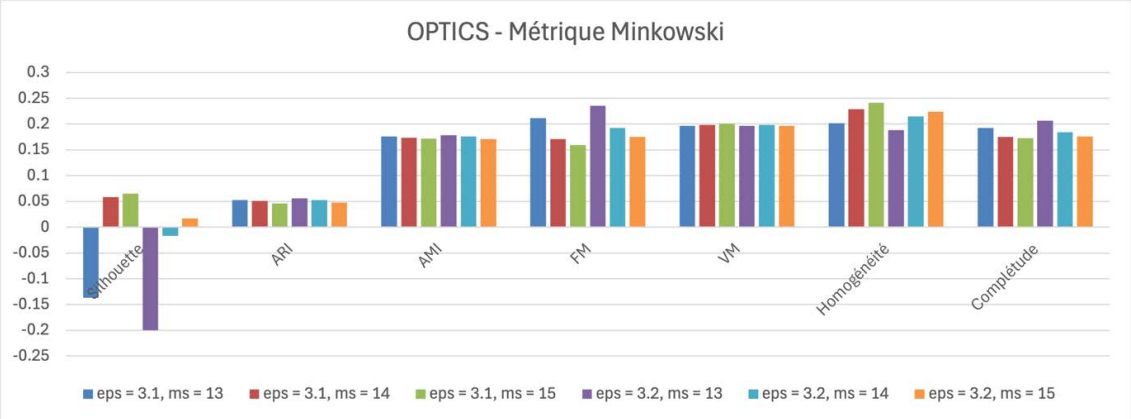


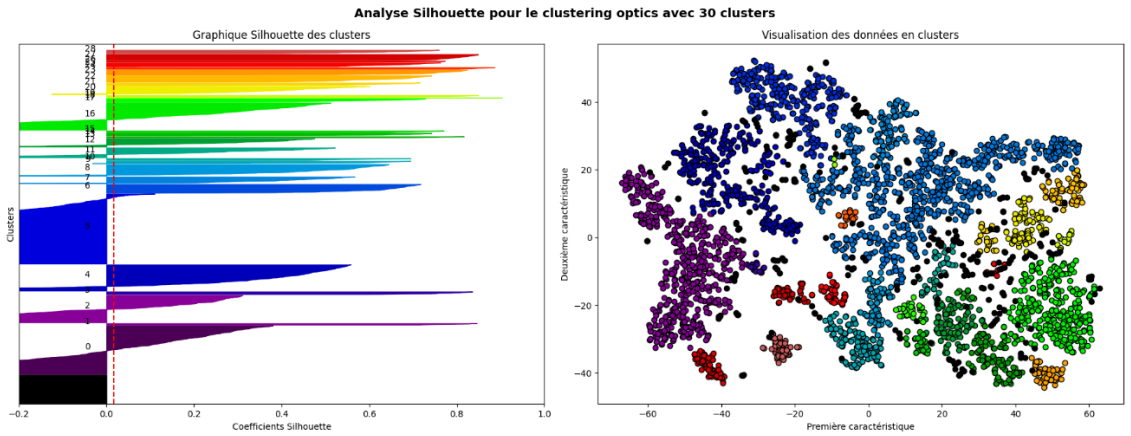
Figure 5.33 Graphique en barres des métriques d'évaluation des clusters OPTICS utilisant la métrique de distance Minkowski, avec différents paramètres

OPTICS – Métrique Minkowski

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

Paramètres : *eps = 3,2, min_cluster_size = 5, min_samples = 15*

Résultat : 30 clusters



a *b*

Figure 5.34 Clustering OPTICS utilisant la métrique de distance Minkowski: graphique Silhouette (a) et graphique 2D en nuage de points (b)

En comparant les graphiques en nuages de points de la Figure 5.32 (distance Euclidienne) et de la Figure 5.34 (distance Minkowski), nous remarquons peu de différences dans l'apparence des clusters formés. De plus, comme mentionné précédemment, en observant les images 2D des échantillons, nous n'observons pas d'amélioration majeure dans la répartition des échantillons parmi les clusters. Nous avons donc choisi d'afficher les images 2D des cinq premiers échantillons appartenant à chacun des 30 clusters obtenus avec OPTICS utilisant la métrique de distance Minkowski, puisque c'est la métrique par défaut de cet algorithme. Nous avons aussi affiché les images 2D des 15 premiers échantillons identifiés comme étant du bruit. Ces images sont présentées en annexe (ANNEXE VIII, Figure-A VIII-1 et Figure-A VIII-2).

Finalement, nous avons appliqué l'algorithme OPTICS sur les **données non étiquetées**. Nous avons utilisé les paramètres : métrique de distance = Minkowski, *min_cluster_size* = 5 et *min_samples* = 15, avec lesquels nous avons obtenu de bons résultats sur les données étiquetées, mais nous avons modifié le paramètre *eps* à 3,0. La Figure 5.35 présente le

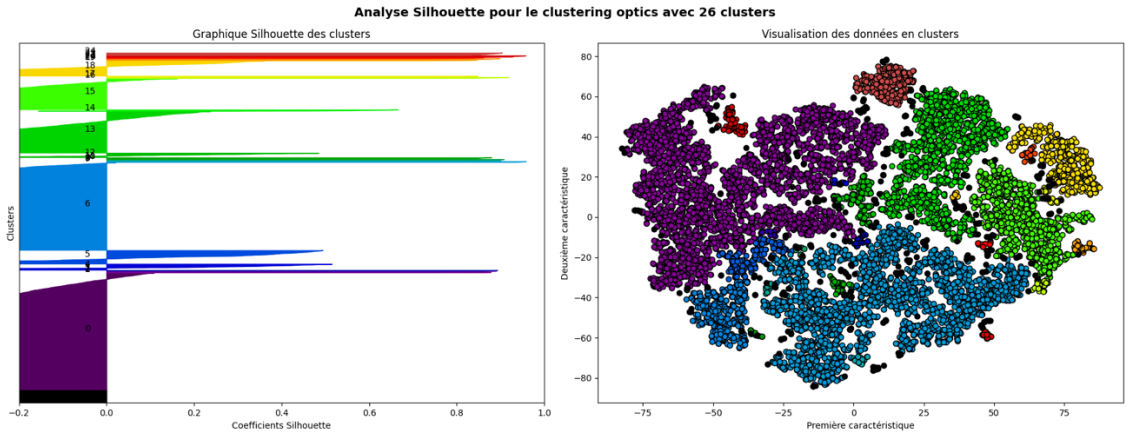
graphique Silhouette ainsi que le graphique en nuages de points pour cette expérimentation faite à partir des caractéristiques extraites par ENB3 et réduites à 2 dimensions par TSNE.

OPTICS – Distance Minkowski

Données non étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

Paramètres : *eps = 3,0; min_cluster_size = 5; min_samples = 15*

Résultat : 26 clusters



a *b*

Figure 5.35 Clustering OPTICS utilisant la métrique de distance Minkowski avec les données non étiquetées : graphique Silhouette (a) et graphique 2D en nuage de points (b)

5.3.2.6 Spectral

Le clustering spectral est le dernier algorithme de clustering traditionnel évalué dans ce chapitre. Avec les paramètres, plusieurs configurations sont possibles. Nous avons donc débuté avec une analyse préliminaire en effectuant une recherche manuelle et en observant les graphiques en nuages de points des clusters, afin d'identifier une plage de paramètres intéressants. Tout comme nous l'avons fait pour les algorithmes précédents, nous avons cherché des graphiques illustrant 15 à 30 clusters, bien définis et avec des frontières claires. D'après ces itérations préliminaires, dont les résultats ne sont pas présentés dans cette section, nous avons choisi de présenter les expérimentations variant les paramètres *affinity* et *n_clusters*, en utilisant la méthode cluster_qr pour *assign_labels*. Pour les autres paramètres, nous avons choisi d'utiliser les valeurs par défaut, notamment *gamma* = 1,0;

eigen_solver = arpack et *n_components = None* (dans ce cas, l'algorithme utilise la valeur établie pour *n_clusters*). Les paramètres sélectionnés sont indiqués dans le Tableau 5.6.

Dans cette section, nous présentons les résultats des expérimentations effectuées à partir des caractéristiques extraites par le modèle ENB3 et réduites à 2 dimensions par TSNE. Nous avons aussi tenté d'appliquer le clustering spectral sans réduire la dimensionnalité des caractéristiques, étant donné que l'algorithme effectue aussi cette tâche. Pour ce faire, nous avons utilisé les caractéristiques extraites normalisées (dim = 32). La qualité des clusters produits était faible et nous avons abandonné cette voie.

Tableau 5.6 Paramètres pour l'algorithme spectral

Paramètre	**Valeurs**
affinity	laplacian, chi2, sigmoid, poly
gamma	1,0
eigen_solver	arpack
n_components	None
assign_labels	cluster_qr
n_clusters	15, 30

La Figure 5.36 présente les métriques calculées pour les itérations utilisant différentes valeurs pour les paramètres *affinity* et *n_clusters*. Les résultats produits par la méthode sigmoïde nous semblent les moins intéressants, car presque toutes les métriques sont inférieures. De plus, en observant le contenu des différents clusters produits, nous remarquons que les noyaux sigmoïdes et poly produisent seulement sept clusters contenant plus d'un échantillon, ce qui nous semble insuffisant pour notre problématique. Nous avons donc choisi de présenter les graphiques Silhouette et les graphiques en nuages de points pour les itérations avec le noyau Laplacien et le noyau CHI2 pour 15 clusters (Figure 5.37) ainsi que pour 30 clusters (Figure 5.38).

Spectral

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

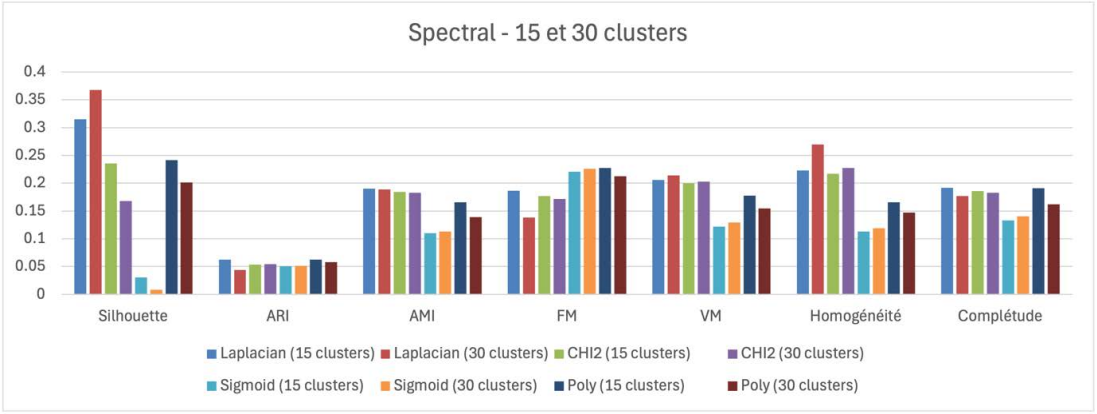


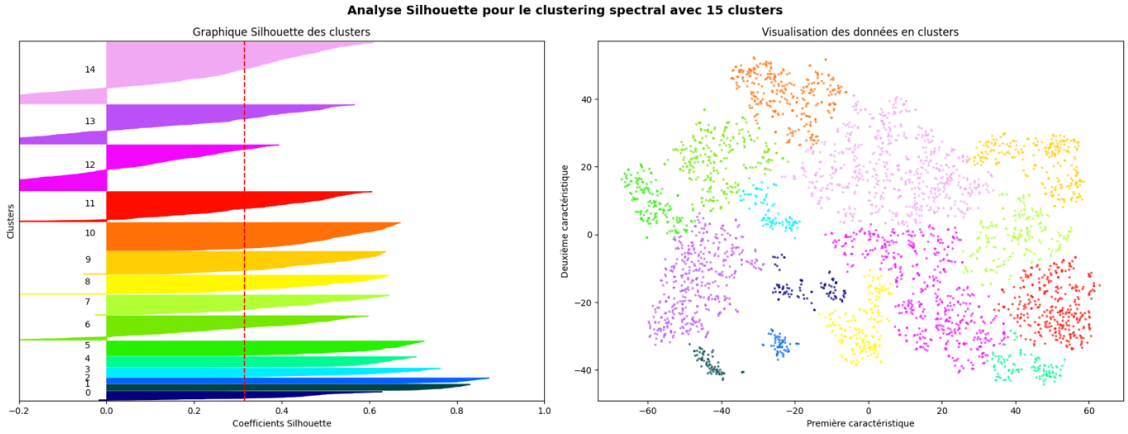
Figure 5.36 Graphique en barres des métriques d'évaluation du clustering spectral utilisant différents paramètres pour calculer la matrice de proximité, pour 15 et 30 clusters

Nous aimons particulièrement la forme des clusters pour le clustering spectral de 30 clusters avec le noyau CHI2, car il présente des clusters de tailles variées, ainsi que quelques clusters ne contenant qu'un seul élément, rappelant la distribution déséquilibrée de nos catégories. Nous avons donc affiché les images 2D de l'échantillon contenu dans chacun des 13 premiers clusters, ne contenant qu'un seul échantillon, ainsi que les cinq premiers échantillons de chacun des autres clusters en annexe (ANNEXE IX, Figure-A IX-1 et Figure-A IX-2).

Spectral

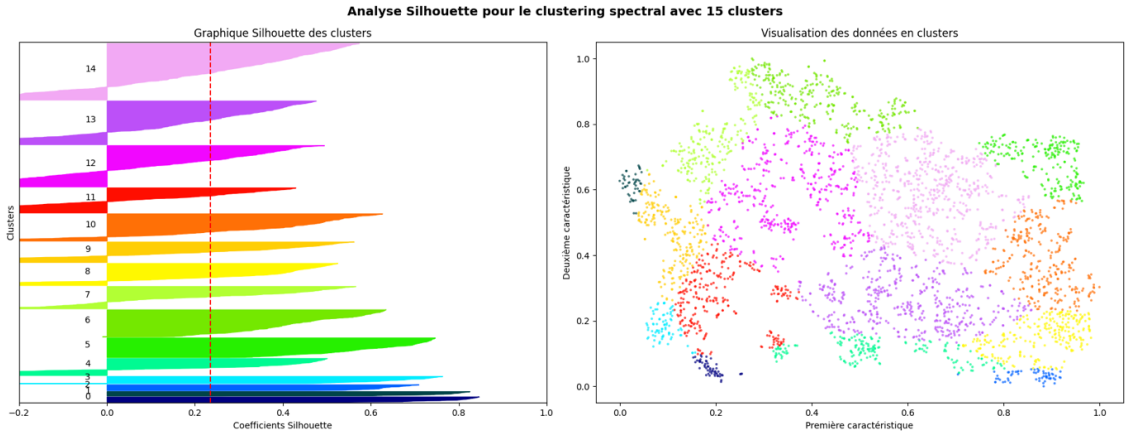
Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

Paramètres : *affinity = Laplacian, assign_labels = cluster_qr, n_clusters = 15*



a *b*

Paramètres : *affinity = CHI2, assign_labels = cluster_qr, n_clusters = 15*



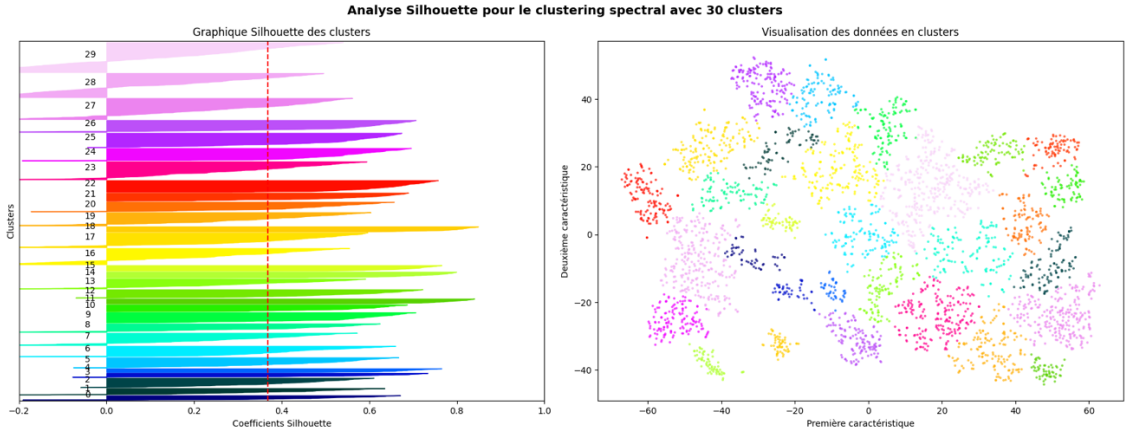
c *d*

Figure 5.37 Clustering spectral de 15 clusters avec noyau laplacien et CHI2 : graphiques Silhouette (a, c) et graphiques 2D en nuage de points (b, d)

Spectral

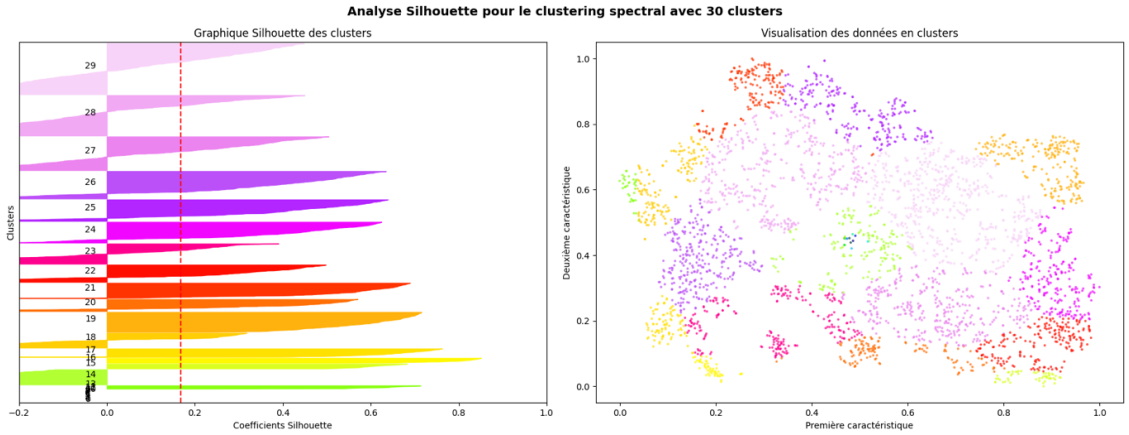
Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

Paramètres : *affinity = Laplacian, assign_labels = cluster_qr, n_clusters = 30*



a *b*

Paramètres : *affinity = CHI2, assign_labels = cluster_qr, n_clusters = 30*



c *d*

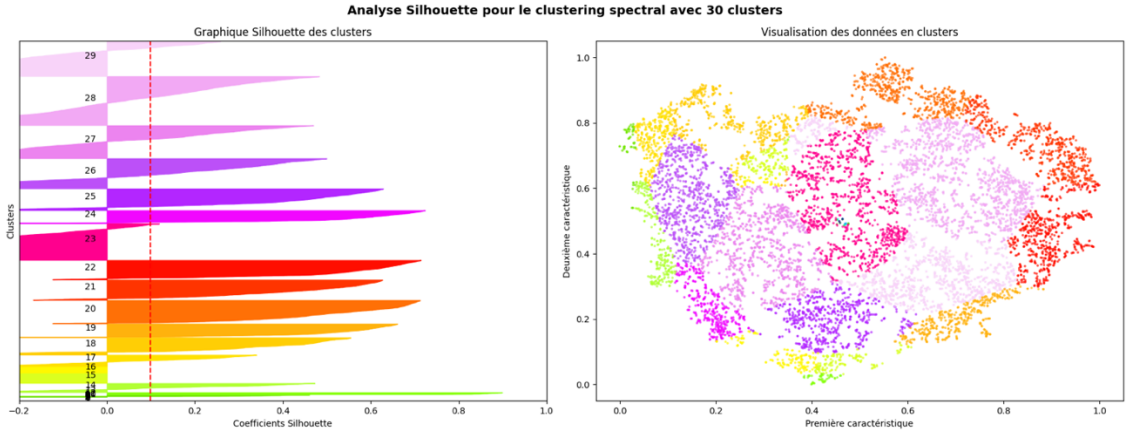
Figure 5.38 Clustering spectral de 30 clusters avec noyau laplacien et CHI2 : graphiques Silhouette (a, c) et graphiques 2D en nuage de points (b, d)

Finalement, nous avons appliqué l'algorithme spectral sur les **données non étiquetées**. Nous avons utilisé les paramètres *affinity* = CHI2, *assign_labels* = cluster_qr et *n_clusters* = 30, avec lesquels nous avons obtenu de bons résultats sur les données étiquetées. La Figure 5.39 présente le graphique Silhouette ainsi que le graphique en nuages de points pour cette expérimentation faite à partir des caractéristiques extraites par ENB3 et réduites à 2 dimensions par TSNE.

Spectral

Données non étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

Paramètres : *affinity = CHI2, assign_labels = cluster_qr, n_clusters = 30*



a *b*

Figure 5.39 Clustering spectral de 30 clusters utilisant le noyau CHI2, avec les données non étiquetées : graphique Silhouette (a) et graphique 2D en nuage de points (b)

5.3.3 Regroupement par réseau profond de clustering

Pour ces expérimentations, nous avons débuté avec les deux ensembles utilisés précédemment pour entraîner les autoencodeurs : AE-I, les images étiquetées avec le bruit gaussien et AE-II, les images étiquetées sans bruit gaussien. Avant de fournir les images au réseau DCN, nous les avons redimensionnées à 100x100, puis nous les avons modifiées pour qu'elles aient une forme de 1x10000. Nous avons ensuite appliqué une mise à l'échelle et de la normalisation sur les vecteurs. Nous avons aussi tenté quelques expérimentations avec des images redimensionnées à 300x300, mais les temps d'entraînement étaient beaucoup trop élevés pour les gains observés.

L'entraînement du réseau DCN s'exécute en deux étapes. D'abord, une étape de préentraînement permet d'initialiser les paramètres de l'autoencodeur. Puis, la seconde étape intègre le clustering à l'entraînement de l'autoencodeur. Plusieurs paramètres peuvent être configurés pour le DCN. D'abord, le nombre et la dimension des couches pour l'encodeur (le décodeur sera construit comme un miroir de l'encodeur), ainsi que la dimension de l'espace

latent. Nous avons testé différentes configurations : par exemple [4096, 1024, 256, 128, 32] avec un espace latent de 2; et [2048, 1024, 256] avec un espace latent de 2. Ensuite, le paramètre *clust_loss_weight*, correspondant au terme $\frac{\lambda}{2}$ de l'équation (5.1), doit être optimisé. Nous avons testé quelques valeurs, variant de 0,1 à 100. Aussi, puisque le DCN utilise l'algorithme de clustering K-means, le nombre de clusters doit être spécifié. Mais il s'agit seulement d'une indication de départ, car ce nombre sera recalculé par le réseau à chaque époch lors de l'entraînement. Nous avons utilisé une valeur de départ de 30. Le Tableau 5.7 résume les paramètres que nous avons testés, et le Tableau 5.8 présente les paramètres d'entraînement, tels que la taille des lots (batch size), le nombre d'époch pour le préentraînement ainsi que pour l'entraînement, et le taux d'apprentissage (learning rate).

Tableau 5.7 Paramètres de configuration du DCN

Paramètre	**Valeurs**
***Couches* → Espace latent**	[4096, 1024, 256, 128, 32] → 2 [2048, 1024, 256] → 2
clust_loss_weight	0,1; 0,5; 1; 2; 5; 100
Nb_clusters	30
Taille de l'image d'entrée	300x300; 100x100

Tableau 5.8 Paramètres d'entraînement du DCN

Paramètre	**Valeurs**
Taille du lot	64
Nb epochs préentraînement	3; 5; 10
Nb epochs entraînement	3; 10; 50
Taux d'apprentissage	0,0005; 0,05; 0,5

Malgré toutes les configurations testées, nous ne sommes pas parvenus à obtenir des résultats satisfaisants. D'abord, nous remarquons que peu importe les paramètres utilisés, le clustering semble s'effectuer sur la forme générale de la région d'intérêt (le grand et le petit cercle),

plutôt que sur les marques microscopiques que nous recherchons. Cette observation est d'autant plus évidente lorsque le DCN est appliqué sur les données étiquetées sans bruit gaussien, car les zones noires sont plus faciles à comparer visuellement. Les images de la Figure 5.40 montrent cinq images de quelques clusters pour l'une de ces expérimentations, utilisant les paramètres suivants : tailles des couches de l'autoencodeur = [2048, 1024, 256], taille de l'espace latent = 2, *clust_loss_weight* = 2, nb épochs pour le préentraînement = 5, nb épochs pour l'entraînement = 10 et taux d'apprentissage = 0,0005.

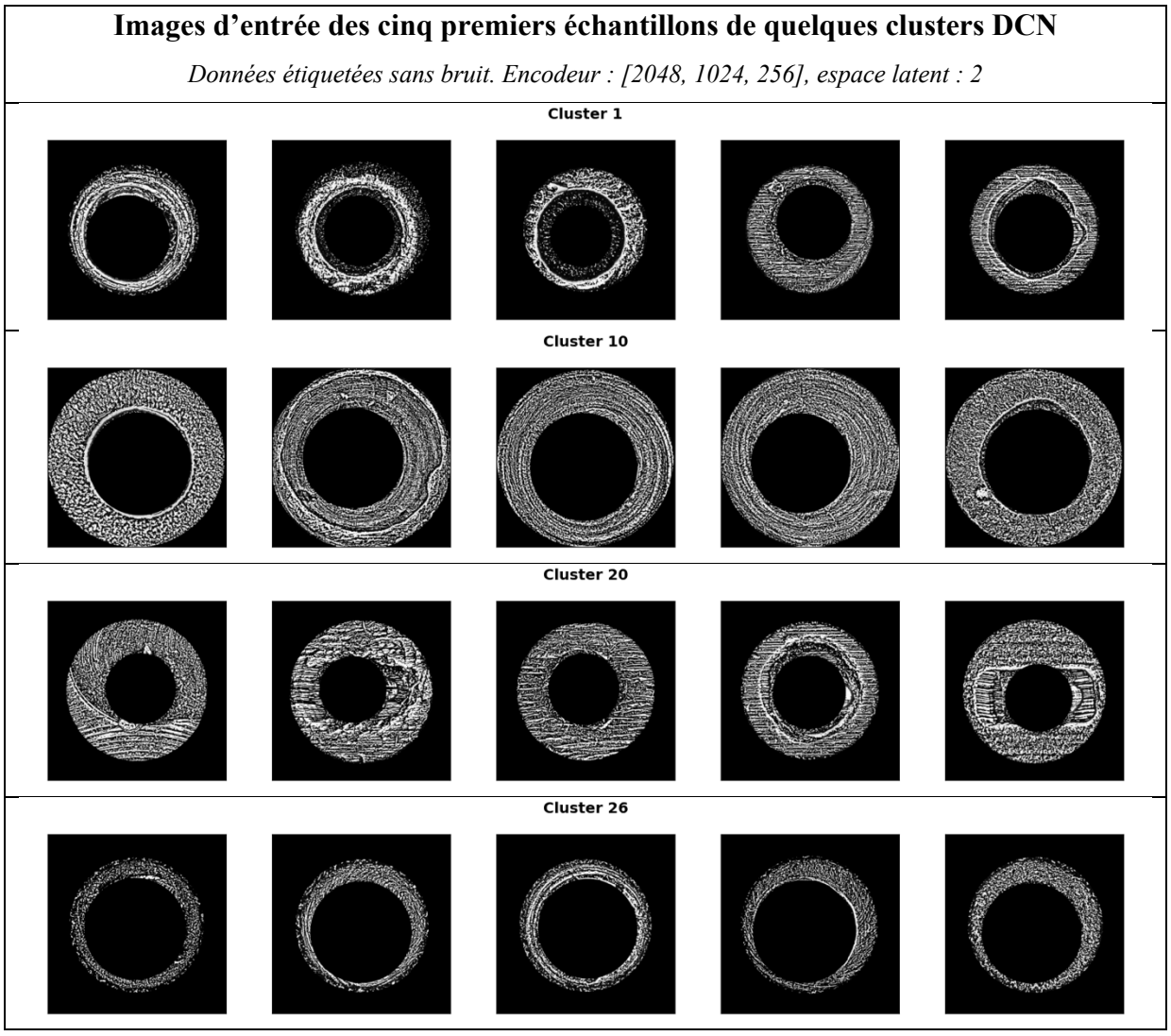


Figure 5.40 Images d'entrée des cinq premiers échantillons de quelques clusters DCN

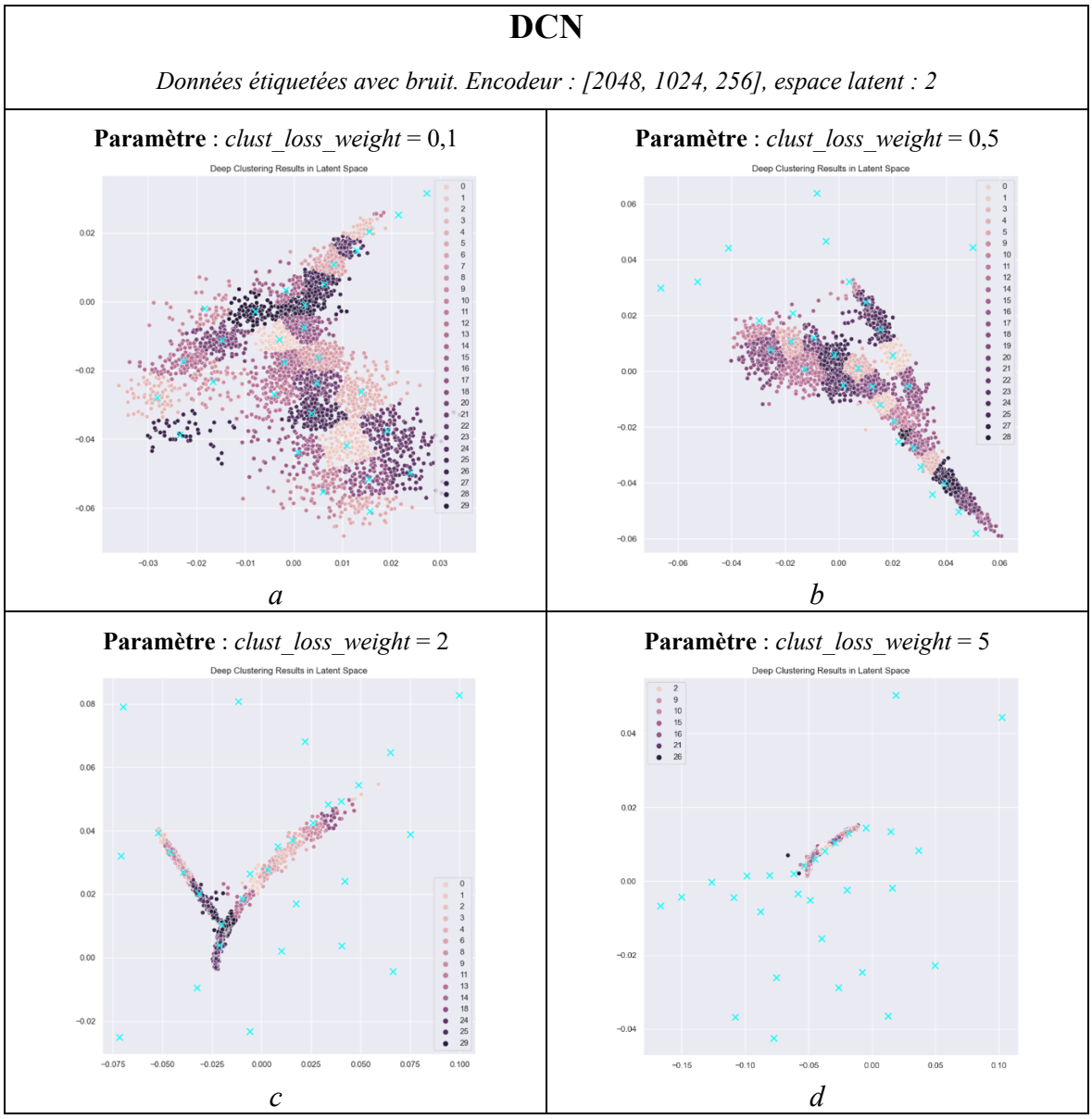


Figure 5.41 Clustering DCN pour différentes valeurs du paramètre *clust_loss_weight* : graphiques 2D en nuage de points pour clw = 0,1 (a), 0,5 (b), 2 (c), et 5 (d)

En ce qui concerne l'entraînement du réseau, nous avons remarqué que la perte ne change pratiquement pas lors de l'étape de préentraînement. Ainsi, il semble inutile d'effectuer cette tâche sur plusieurs époch. Par suite de ces observations, nous avons évalué de nouveau les résultats obtenus dans la section 5.3.2.1, lorsque nous avons testé le clustering K-means sur les caractéristiques extraites par des autoencodeurs. Nous remarquons un comportement similaire, mais moins prononcé. Toutefois, ce comportement explique les résultats décevants

que nous avions remarqués lorsque nous avons utilisé les autoencodeurs. Ainsi, les expérimentations menées jusqu'à présent nous indiquent que, lorsque les autoencodeurs sont utilisés pour extraire les caractéristiques, ils s'attardent avant tout aux formes générales et négligent les marques microscopiques qui nous intéressent.

De plus, l'augmentation du paramètre *clust_loss_weight* ne semble pas améliorer les résultats. Plus nous augmentons la valeur de ce paramètre et plus les échantillons se rapprochent dans l'espace latent, réduisant la distance entre les clusters ainsi que leur nombre, ce qui nous paraît indésirable. Les images de la Figure 5.41 montrent des graphiques en nuages de points illustrant cette réduction des clusters. Les graphiques ont été obtenus en variant le paramètre *clust_loss_weight*, avec les autres paramètres fixés aux valeurs suivantes : tailles des couches de l'autoencodeur = [2048, 1024, 256], taille de l'espace latent = 2, nb épochs pour le préentraînement = 3, nb épochs pour l'entraînement = 3 et taux d'apprentissage = 0,0005.

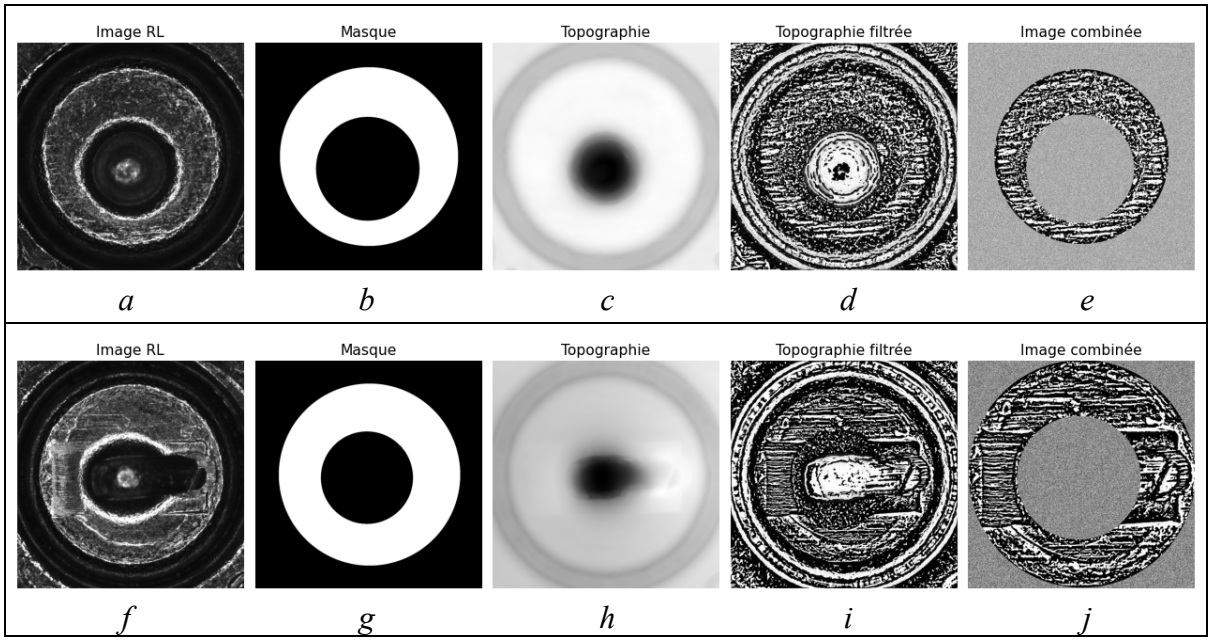


Figure 5.42 Traitement d'une image topographique. Image originale 2D (a, f); masque correspondant (b, g); image topographique (c, h); topographie filtrée (d, i); topographie filtrée, sur laquelle le masque binaire a été superposé et remplacé par un bruit gaussien aléatoire (e); puis rognée aux bordures de la région d'intérêt (j)

Puisqu'il nous semble inutile de présenter les métriques obtenues pour ces dernières expérimentations, nous avons plutôt tenté d'améliorer la performance du réseau DCN en modifiant les images d'entrée. Nous avons ainsi créé un nouvel ensemble : pour chaque image, nous avons utilisé les dimensions du cercle externe afin de rogner l'image aux limites de la région d'intérêt. Cette opération permet d'uniformiser la taille des cercles extérieurs; toutefois, les cercles intérieurs demeurent de tailles variées. La Figure 5.42 (j) présente un exemple de cette nouvelle configuration des images pour l'entrée du DCN (rognées). À titre de rappel, la Figure 5.42 (e) présente l'ancienne configuration (bruit). De plus, nous avons abandonné l'idée d'un ensemble sans bruit gaussien, car les zones noires sont les premières reconnues par les encodeurs.

DCN

Données étiquetées avec bruit. Encodeur : [2048, 1024, 256], espace latent : 2

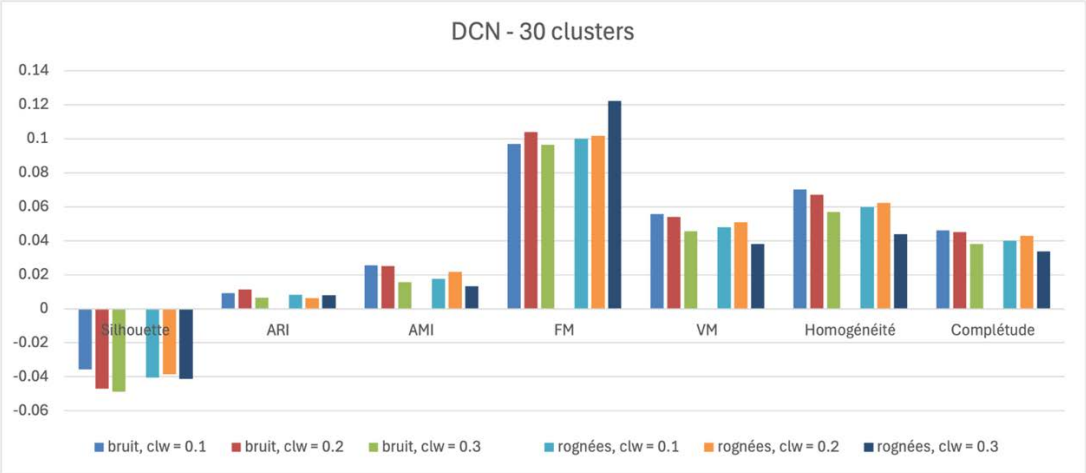


Figure 5.43 Graphique en barres des métriques d'évaluation des clusters DCN avec variation du paramètre *clust_loss_weight*, pour 30 clusters

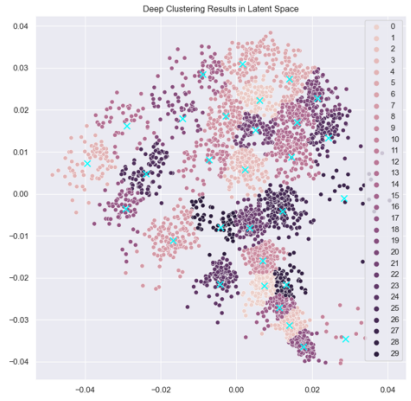
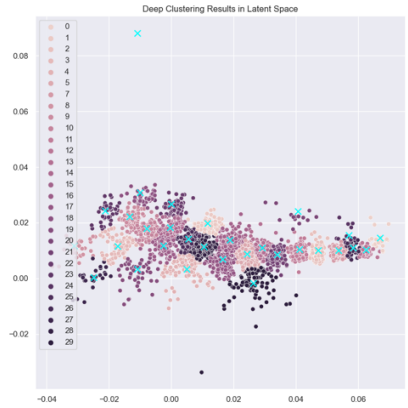
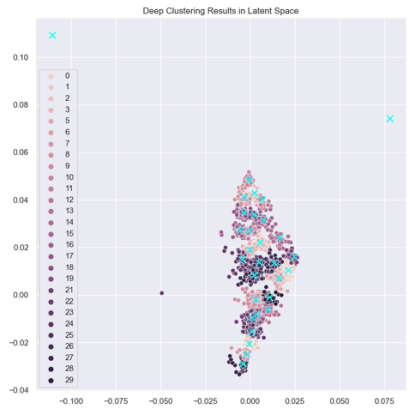
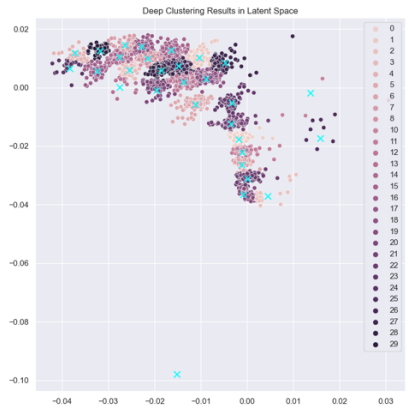
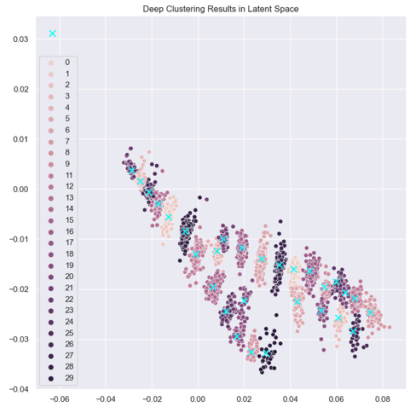
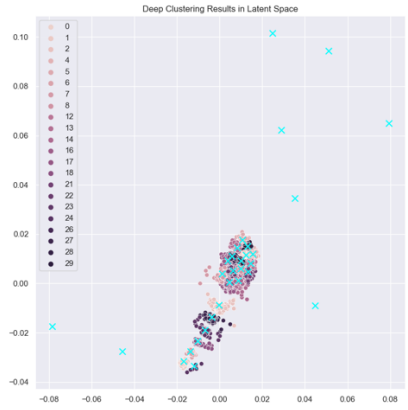
DCN *Encodeur : [2048, 1024, 256], espace latent : 2. Nb époch préentraînement : 5, entraînement : 40*		
clust_loss_weight	**Données étiquetées avec bruit**	**Données étiquetées rognées**
0,1	 *a*	 *b*
0,2	 *c*	 *d*
0,3	 *e*	 *f*

Figure 5.44 Clustering DCN pour différentes valeurs du paramètre *clust_loss_weight* : graphiques 2D en nuage de points pour *clw* = 0,1 (a, b), 0,2 (c, d), 0,3 (e, f)

Nous avons repris les entraînements en utilisant les paramètres suivants : tailles des couches de l'autoencodeur = [2048, 1024, 256], taille de l'espace latent = 2, nb épochs pour le préentraînement = 5, nb épochs pour l'entraînement = 40 et taux d'apprentissage = 0,0005. La Figure 5.43 présente les métriques calculées pour les itérations de variation du paramètre *clust_loss_weight* avec les valeurs 0,1; 0,2 et 0,3. Les images de la Figure 5.44 montrent des graphiques en nuages de points pour ces expérimentations.

Nous remarquons que les graphiques produits par les expérimentations utilisant les images étiquetées avec bruit (Figure 5.44 a, c, e) semblent former des groupes plus distincts les uns des autres. De plus, les métriques présentées sont en général plus élevées pour les expérimentations avec ces images (sauf pour le coefficient Silhouette moyen). Cependant, en observant les images 2D des échantillons contenus dans les différents clusters, nous ne remarquons pas de différences significatives. Malheureusement dans tous les cas, le réseau semble toujours prioriser les éléments qui nous intéressent moins, tels que la forme de la région d'intérêt. Nous avons affiché en annexe les images 2D des cinq premiers échantillons de chacun des 28 clusters formés lors de l'expérimentation avec le paramètre *clust_loss_weight* = 0,2, utilisant les images étiquetées avec bruit (ANNEXE X, Figure-A X-1). Nous croyons que du travail supplémentaire serait nécessaire afin de concevoir des autoencodeurs plus performants, et capables de reconstituer correctement la texture de l'image; et subséquemment modifier les encodeurs du DCN en conséquence.

5.4 Conclusion

Afin de faciliter les comparaisons, un sommaire des paramètres utilisés pour chacun des algorithmes de clustering est disponible en annexe (ANNEXE XI, Tableau-A XI-1). De plus, la Figure-A XI-1 permet de visualiser les graphiques en nuages de points obtenus pour les algorithmes testés sur toutes les données non étiquetées. Les résultats pour les algorithmes de clustering traditionnels ont été obtenus à partir des caractéristiques extraites par la CNN ENB3 et réduites à 2 dimensions par TSNE.

Nous sommes d'avis que l'évaluation des clusters est une tâche ardue. Dans notre cas, nous avons rencontré plusieurs difficultés. Tout d'abord, les métriques obtenues grâce aux étiquettes des échantillons n'ont pas été particulièrement représentatives des résultats. En partie à cause de la présence d'un nombre élevé d'échantillons étiquetés comme étant « inconnu », mais probablement aussi à cause des incohérences de la vérité de terrain qui ont été discutées dans le chapitre précédent. Nous les avons toutefois utilisées, principalement pour guider les sélections de paramètres. Mais nous n'avons pas trouvé qu'elles représentaient un bon indicateur pour l'évaluation de la conformité des éléments contenus dans les clusters, ou pour comparer les différents algorithmes.

Le coefficient Silhouette, qui ne dépend pas de la vérité de terrain, n'a pas été bien plus utile. Nous l'avons principalement utilisé, accompagné d'un graphique en nuages de points, pour tenter d'interpréter le comportement des regroupements. Ces renseignements ont permis, encore une fois, de guider nos choix de paramètres. Cependant, cette évaluation de la qualité des clusters s'est avérée négligeable pour qualifier l'adéquation des résultats. En outre, les algorithmes de clustering K-means et Fuzzy C-means ont obtenus des valeurs de coefficient Silhouette plus élevées, car cette métrique est mieux adaptée pour les algorithmes formant des clusters sphériques de tailles similaires. Lorsque nous l'avons utilisé pour les algorithmes se basant sur la densité, tels que DBSCAN, HDBSCAN et OPTICS, nous avons réalisé que cette métrique n'est pas idéale lorsque les formes et les tailles des clusters varient, et encore moins avec la présence de valeurs aberrantes, qui sont placées un peu partout et réduisent davantage la valeur de la métrique. De façon générale, l'observation des images 2D associées aux échantillons dans les différents clusters a pu nous fournir quelques indications. Mais des évaluations par des experts seraient nécessaires pour pouvoir avancer plus loin. Nous avons aussi rencontré des problèmes au niveau de l'extraction des caractéristiques. D'abord, les modèles préentraînés que nous avons utilisés sont affligés des problèmes discutés dans le chapitre précédent. Nous nous questionnons sur le fait qu'ils soient capables d'extraire des caractéristiques pertinentes pour toutes les catégories de marques. Des expérimentations supplémentaires utilisant des CNN entraînées sur des données étiquetées vérifiées seraient souhaitables.

Du côté des autoencodeurs, nous avons réalisé que les modèles de base ne parviennent pas à cartographier correctement les informations de texture dans l'espace latent. Nous observons cela à travers la reconstruction des images, qui se concentre principalement sur les formes générales de la région d'intérêt, et appliquent une zone floue aux emplacements où il devrait y avoir les marques que nous cherchons à identifier (voir les exemples de la Figure 5.5). Ce problème a aussi été observé avec le clustering par DCN, qui utilise des autoencodeurs afin d'extraire les caractéristiques. Des modèles d'autoencodeurs plus complexes seraient probablement nécessaires pour cette tâche. Bref, des travaux futurs visant à améliorer la détection de la texture fine contenue dans la région de la marque de la culasse pourraient certainement contribuer à améliorer les résultats obtenus. Nous avons observé que les algorithmes favorisant la création de clusters de formes et de tailles variées semblent mieux adaptés à la représentation de nos données déséquilibrées. Nous avons également apprécié la technique de clustering flou, utilisée par l'algorithme Fuzzy C-Means, qui non seulement assigne un cluster pour chaque point de l'ensemble, mais calcule aussi un degré d'appartenance de chaque point à chaque groupe. En déterminant une valeur seuil pour ces degrés d'appartenance, il pourrait être possible d'attribuer les catégories multiétiquettes de chaque point, en utilisant un nombre de clusters égal au nombre de catégories de marques (six, ou sept en incluant la catégorie « inconnu »). En dernière analyse, il ressort de cela que des travaux futurs pourraient tenter d'utiliser une technique de clustering flou avec un algorithme permettant la formation de clusters de tailles variables, tel que le clustering spectral, afin d'obtenir les degrés d'appartenance de chaque échantillon.

Finalement, nous avons identifié une autre piste de recherche pour l'utilisation du clustering. Plutôt que d'appliquer le clustering sur les caractéristiques descriptives de l'ensemble des images, il pourrait être utilisé sur chaque image séparément pour regrouper les pixels similaires dans une image afin d'en faire une segmentation des différents éléments, qui pourraient ensuite être classifiés. Ces travaux pourraient être considérés dans le futur.

CHAPITRE 6

CONTRIBUTION #3 – ACCORD INTEROBSERVATEUR ENTRE OBSERVATEURS HUMAINS ET MÉTHODES D'APPRENTISSAGE

6.1 Introduction

Ce chapitre discute des expérimentations pour réaliser le dernier objectif de cette thèse, soit l'étude de l'accord interobservateur entre des observateurs humains et les méthodes d'apprentissage profond supervisé. Il débute avec une explication des préalables, suivie d'une présentation de la méthodologie incluant le détail de l'implémentation. Le chapitre se poursuit avec une présentation des résultats pour les différentes expérimentations, accompagnée d'observations.

Cet objectif a été élaboré à la suite des expérimentations précédentes, au cours desquelles nous avons observé des incohérences dans l'ensemble de vérité de terrain initial (la vérité de terrain que nous avons utilisé jusqu'à maintenant). Nous croyons que ces incohérences auraient pu affecter négativement les résultats obtenus, en entravant l'apprentissage de caractéristiques descriptives pertinentes lors des entraînements supervisés des modèles. De surcroît, nous supposons que ces entraînements imparfaits auraient pu affecter négativement les résultats obtenus lors des expérimentations non supervisées, car nous avons utilisé ces modèles préentraînés afin d'extraire les caractéristiques descriptives des images. Ainsi, nous avons déterminé cet objectif afin, d'une part, de mieux comprendre la difficulté de la tâche d'annotation des échantillons en comparant les étiquettes obtenues par différents observateurs, et d'autre part, d'évaluer si un ensemble de vérité de terrain vérifié pourrait améliorer les résultats obtenus jusqu'à maintenant.

6.2 Préalables

6.2.1 Demande d'approbation au Comité d'éthique de la recherche de l'ÉTS

Puisque la participation d'experts humains fut nécessaire afin d'effectuer l'annotation des échantillons, nous avons débuté avec la préparation d'une demande d'approbation adressée au comité d'éthique de la recherche de l'ÉTS. Entre autres, cette demande comprend une description sommaire du projet, ainsi que le détail des tâches demandées aux observateurs humains. Une fois l'approbation du comité d'éthique obtenue, nous avons poursuivi avec la partie concernant l'accord interobservateur, et la partie concernant l'ensemble de vérité de terrain vérifié.

6.2.2 Outil logiciel pour l'annotation manuelle

Afin de faciliter le travail des annotateurs, nous avons implémenté un outil logiciel permettant de parcourir une sélection d'échantillons. Un bouton de l'interface (Figure 6.1) permet d'afficher différentes images pour la visualisation de chaque échantillon. Ainsi, il est possible de faire défiler les images 2D avec un éclairage annulaire et un éclairage de côté, ainsi qu'une reconstruction 2D de la topographie 3D. Toutefois, l'outil n'inclut pas de fonctionnalités de visualisation avancées, telles que le déplacement ou la rotation des images, ou la modification de l'éclairage. Des cases à cocher permettent ensuite de sélectionner les catégories observées. Comme l'annotation est multiétiquette, il est possible de sélectionner plus d'une case, à l'exception de la case « inconnu » qui doit être sélectionnée seule. Si cette dernière règle n'est pas suivie, un message avise l'utilisateur et l'annotation devra être corrigée afin de pouvoir poursuivre ou de sauvegarder le travail. Des boutons de navigation permettent de parcourir les échantillons vers l'avant, ou de revenir en arrière pour corriger les annotations d'un échantillon précédent. Un bouton de sauvegarde permet d'interrompre le travail et de le reprendre plus tard. Un champ de texte libre optionnel permet d'ajouter des notes, selon le jugement de l'annotateur. Par exemple, elles pourraient être utilisées pour clarifier un choix difficile. Les annotations sont sauvegardées dans un fichier .csv, que nous pouvons ensuite récupérer afin d'évaluer les résultats et calculer les métriques. Nous avons

installé cet outil informatique sur un serveur, afin de permettre aux utilisateurs de se connecter à partir d'un poste de travail de leur choix, et d'effectuer le travail à distance.

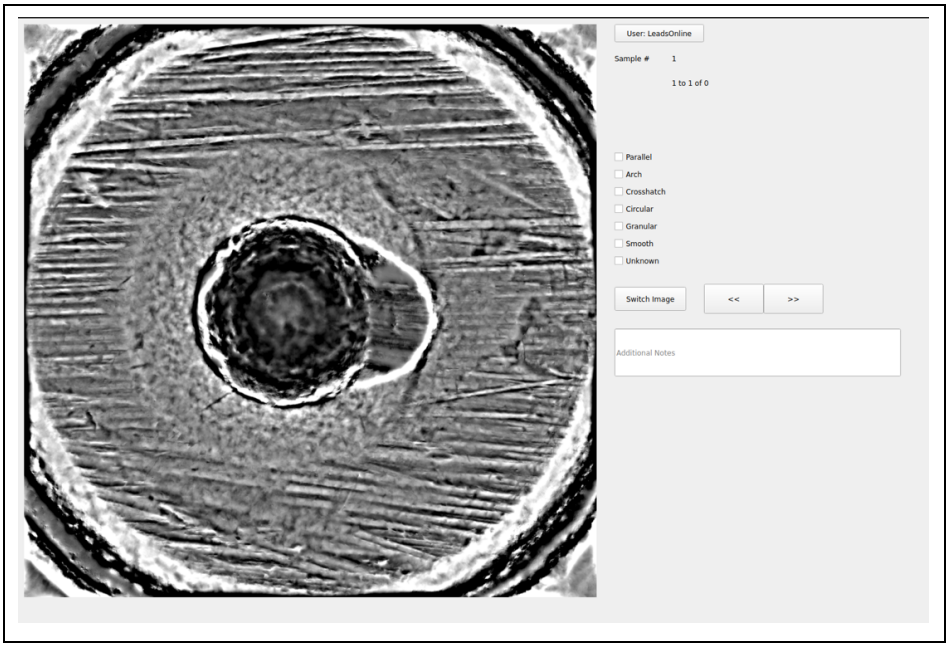


Figure 6.1 Interface principale de l'outil informatique

6.3 Méthodes

Cette section détaille la méthodologie suivie pour cet objectif. Nous avons divisé ce chapitre en trois parties. La première partie tente d'établir des métriques d'accord interobservateur, calculées à partir des étiquettes d'un ensemble de 100 échantillons, obtenues par l'annotation de six observateurs humains, les prédictions de trois observateurs machines, ainsi que la vérité de terrain initiale (VT-i). Nous avons choisi d'utiliser le coefficient Kappa afin de calculer ces métriques. La seconde partie se charge de créer un ensemble de vérité de terrain vérifié (VT-v), contenant 1000 échantillons, dont les étiquettes ont été approuvées par deux experts. Finalement, la troisième partie reprend les expérimentations de classification multiétiquettes effectuées au cours du premier objectif, en utilisant ce nouvel ensemble de vérité de terrain vérifié. Nous y présentons les résultats obtenus, et nous les comparons à ceux obtenus avec l'ensemble de vérité de terrain initial, présentés au CHAPITRE 4. Pour ce chapitre, nous avons utilisé des échantillons provenant du BrassTRAX HW4 (section 3.3.2).

6.3.1 Implémentation de la première partie : l'accord interobservateur

Cette section présente l'implémentation de la première partie de cette contribution : l'accord interobservateur. Le terme observateur a été choisi afin d'inclure des observateurs humains (les annotateurs), des observateurs machines (les modèles), ainsi que la vérité de terrain initiale comme un observateur additionnel.

Pour cette partie, nous avons sélectionné 100 échantillons exhibant des marques variées, qui nous semblaient différents les uns des autres. Nous nous sommes vaguement référés à la vérité de terrain initiale pour guider ces choix, en essayant de sélectionner le maximum de configurations de marques possibles. À cette fin, nous avons parcouru tous les échantillons pour trouver des échantillons qui nous semblaient représentatifs.

Ensuite, nous avons recruté six (6) annotateurs, pour tenir le rôle des observateurs humains, que nous avons nommés H-1 à H-6. Puis, nous avons demandé à ces observateurs humains d'utiliser l'outil informatique que nous avons implémenté, chacun de leurs côtés, afin de visionner et d'annoter les 100 échantillons que nous avons sélectionnés. Ce travail, que nous avons estimé nécessiter entre 45 et 90 minutes par annotateur, a pu s'effectuer entièrement à distance, en se connectant au serveur. Une fois les annotations terminées, nous avons récupéré les fichiers d'annotations. Nous avons ensuite sélectionné les trois (3) observateurs machines. Pour chacune des trois architectures évaluées dans le premier objectif, nous avons choisi le meilleur modèle multiétiquette entraîné : VGG16, ENB3 et ViT B32, que nous avons nommés M-1 à M-3. Puis, nous avons utilisé ces modèles afin de prédire les étiquettes des 100 échantillons sélectionnés.

Pour terminer, pour chaque catégorie de marques (parallèle, arche, hachure, circulaire, granulaire, lisse et inconnu), nous avons utilisé le coefficient Kappa pour calculer les métriques d'accord interobservateurs par paires, entre les neuf observateurs (humains et machines), ainsi qu'avec la vérité de terrain initiale (VT-i), que nous considérons ici comme

étant un 10^{e} observateur. Pour chaque observateur, nous avons aussi calculé la somme des échantillons par nombre d'étiquettes, c'est-à-dire le nombre d'échantillons avec une étiquette, deux étiquettes, et ainsi de suite.

6.3.2 Implémentation de la seconde partie : l'ensemble de vérité de terrain vérifié

Pour cette partie, nous avons sélectionné 1000 échantillons, exhibant des marques qui nous semblaient différentes les unes des autres. Tout comme pour la partie précédente, nous avons essayé de sélectionner le maximum de configurations de marques possibles, en parcourant tous les échantillons à la recherche d'exemples qui nous semblaient variés et représentatifs, et en nous guidant vaguement sur la vérité de terrain initiale. Ces 1000 échantillons sont différents des 100 échantillons sélectionnés précédemment pour calculer les accords interobservateurs.

Nous avons ensuite recruté deux (2) annotateurs. Puis, nous leur avons demandé d'utiliser l'outil informatique que nous avons implémenté, afin de visionner les 1000 échantillons que nous avons sélectionnés. Les deux annotateurs devaient observer les images de chaque échantillon et discuter des catégories de marques observées. Ils devaient être en accord avant de saisir les annotations dans l'outil informatique, afin de constituer un seul ensemble de vérité de terrain vérifié. Advenant le cas où un accord est impossible, les instructions précisent que l'échantillon problématique serait éliminé de l'ensemble. Nous avons estimé la durée de ce travail entre 8 et 15 heures. Les annotations pouvaient être effectuées à distance en se connectant au serveur sur lequel l'outil informatique était installé, mais les deux annotateurs devaient pouvoir discuter de chaque échantillon. Nous avons récupéré les fichiers d'annotations une fois le travail complété.

Pour terminer, nous avons utilisé ces annotations pour construire un ensemble de vérité de terrain vérifié, qui a été utilisé pour les expérimentations de la partie suivante. Puis, pour chaque catégorie de marques, nous avons utilisé le coefficient Kappa pour calculer les métriques d'accord interannotateur entre la vérité de terrain vérifiée (VT-v) et la vérité de

terrain initiale (VT-i). Pour chacun de ces deux ensembles de vérité de terrain, nous avons calculé la somme des échantillons par nombre d'étiquette. Finalement, nous avons effectué une analyse exploratoire de ce nouvel ensemble de vérité de terrain.

6.3.3 Implémentation de la troisième partie : classification multiétiquette

Pour cette troisième et dernière partie, nous avons utilisé l'ensemble VT-v afin d'entraîner un modèle de classification. Nous avons d'abord effectué une étape de préparation des données en divisant l'ensemble VT-v en trois sous-ensembles, pour l'entraînement et la vérification des modèles : un ensemble d'entraînement (72%), un ensemble de validation (15%) et un ensemble de test (13%). Lors des expérimentations effectuées pour la première contribution, avec l'ensemble VT-i, nous avons exclu tous les échantillons de la catégorie « inconnu », car leur nombre était beaucoup trop élevé. Pour l'expérimentation multiétiquette avec l'ensemble VT-v, nous avons conservé les échantillons de la catégorie « inconnu », dont le nombre était plus raisonnable.

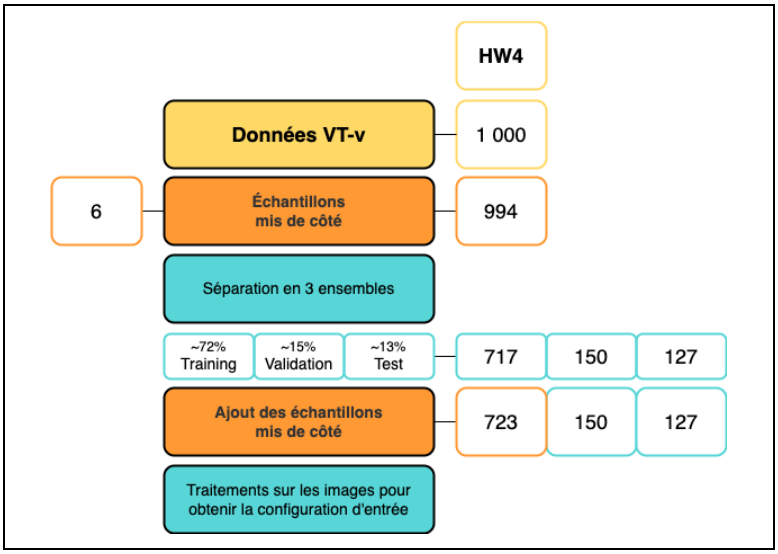


Figure 6.2 Division des données pour l'expérimentation multiétiquette avec VT-v

Comme il sera présenté dans l'analyse exploratoire, nous avons remarqué que six (6) combinaisons de catégories ne contiennent qu'un seul échantillon (Figure 6.4). Pour la

division en sous-ensembles stratifiés, au moins deux échantillons par catégorie sont nécessaires. Puisque nous désirons conserver tous les échantillons de l'ensemble, nous avons mis de côté ces six échantillons avant de faire la division en sous-ensembles. Puis nous les avons ajoutés à l'ensemble d'entraînement. Pour terminer, nous avons appliqué des prétraitements sur les images topographiques afin d'obtenir la configuration d'entrée souhaitée : image topographique filtrée, combinée avec le masque de la région d'intérêt sur lequel un bruit gaussien aléatoire a été ajouté, tel que présenté à la Figure 4.1 (e et j). Le diagramme de la Figure 6.2 illustre les détails de ces étapes.

Lors des expérimentations multiétiquettes présentées dans le premier objectif, nous avons réalisé que l'utilisation de techniques d'augmentation afin de balancer les données d'entraînement, en effectuant un suréchantillonnage des combinaisons de catégories plus rares, n'était pas souhaitable. Particulièrement, cette opération risque de causer un surapprentissage des combinaisons de catégories rares, et suraugmentées. Ainsi, nous avons choisi d'effectuer un entraînement multiétiquette, sans augmentation des données. Nous avons effectué un apprentissage par transfert, en utilisant le modèle EfficientNet B3 préentraîné sur ImageNet, car nous avons observé des performances légèrement supérieures avec ce modèle lors des expérimentations de la première contribution. Nous avons utilisé la tête présentée à la Figure 4.5 et nous avons augmenté la taille de la couche de prédiction à 7, pour inclure la catégorie « inconnu ». Nous avons utilisé une fonction d'activation sigmoïde, jumelée à une perte d'entropie croisée binaire. Le modèle a été compilé avec l'optimiseur Adam, en surveillant la métrique d'exactitude et avec une taille du lot de 32. La fonctionnalité d'arrêt anticipé avec une patience de 25 est toujours utilisée en surveillant l'exactitude de validation. Afin d'évaluer les résultats, nous avons utilisé les métriques utilisées précédemment; c'est-à-dire l'exactitude, la précision, le rappel, la mesure F1, ainsi que les matrices de confusion. Enfin, nous avons entraîné le modèle et nous avons comparé les courbes d'apprentissage pour la performance et l'optimisation, ainsi que les métriques obtenues par les prédictions sur l'ensemble de test, avec les métriques obtenues pour la même expérimentation sur l'ensemble VT-i.

6.4 Résultats

6.4.1 Accord interobservateur

Cette partie présente les résultats des travaux ayant trait aux métriques d'accord interobservateurs. Les tableaux 6.1 à 6.7 présentent les métriques d'accord interobservateurs, pour chacune des sept catégories de marques. Les six observateurs humains sont identifiés H (1 à 6) et les trois observateurs machines sont identifiés par M-1 (VGG16), M-2 (ENB3) et M3 (ViT B32).

Tableau 6.1 Accord interobservateur pour la catégorie : Parallèle

Parallèle										
	H-1	**H-2**	**H-3**	**H-4**	**H-5**	**H-6**	**VT-i**	**M-1**	**M-2**	**M-3**
H-1	1,00	0,63	0,55	0,49	0,50	0,72	0,65	0,47	0,43	0,47
H-2		1,00	0,42	0,53	0,36	0,69	0,43	0,32	0,39	0,51
H-3			1,00	0,29	0,70	0,39	0,65	0,53	0,50	0,41
H-4				1,00	0,26	0,51	0,37	0,22	0,25	0,45
H-5					1,00	0,39	0,64	0,35	0,32	0,36
H-6						1,00	0,44	0,25	0,41	0,49
VT-i							1,00	0,56	0,40	0,35
M-1								1,00	0,43	0,35
M-2									1,00	0,27
M-3										1,00

Tableau 6.2 Accord interobservateur pour la catégorie : Arche

Arche										
	H-1	**H-2**	**H-3**	**H-4**	**H-5**	**H-6**	**VT-i**	**M-1**	**M-2**	**M-3**
H-1	1,00	0,71	0,89	0,89	0,67	0,80	0,41	0,76	0,55	0,39
H-2		1,00	0,71	0,71	0,83	0,73	0,45	0,55	0,59	0,43
H-3			1,00	0,78	0,78	0,69	0,52	0,76	0,55	0,39
H-4				1,00	0,67	0,80	0,41	0,63	0,55	0,39
H-5					1,00	0,80	0,63	0,51	0,55	0,39
H-6						1,00	0,46	0,56	0,47	0,34
VT-i							1,00	0,48	0,36	0,22
M-1								1,00	0,47	0,30
M-2									1,00	0,22
M-3										1,00

Tableau 6.3 Accord interobservateur pour la catégorie : Hachure

Hachure										
	H-1	**H-2**	**H-3**	**H-4**	**H-5**	**H-6**	**VT-i**	**M-1**	**M-2**	**M-3**
H-1	1,00	0,70	0,74	0,48	0,72	0,71	0,66	0,51	0,45	0,35
H-2		1,00	0,65	0,40	0,73	0,72	0,67	0,47	0,44	0,27
H-3			1,00	0,44	0,68	0,76	0,62	0,47	0,48	0,27
H-4				1,00	0,49	0,42	0,34	0,33	0,19	0,25
H-5					1,00	0,74	0,77	0,57	0,56	0,43
H-6						1,00	0,76	0,52	0,48	0,31
VT-i							1,00	0,51	0,54	0,33
M-1								1,00	0,62	0,40
M-2									1,00	0,42
M-3										1,00

Tableau 6.4 Accord interobservateur pour la catégorie : Circulaire

Circulaire										
	H-1	**H-2**	**H-3**	**H-4**	**H-5**	**H-6**	**VT-i**	**M-1**	**M-2**	**M-3**
H-1	1,00	0,85	0,85	0,80	0,93	0,89	0,82	0,70	0,76	0,74
H-2		1,00	0,76	0,79	0,92	0,89	0,73	0,69	0,67	0,66
H-3			1,00	0,70	0,84	0,89	0,81	0,69	0,75	0,73
H-4				1,00	0,79	0,75	0,59	0,53	0,61	0,61
H-5					1,00	0,96	0,81	0,76	0,75	0,73
H-6						1,00	0,85	0,73	0,72	0,77
VT-i							1,00	0,73	0,72	0,70
M-1								1,00	0,75	0,60
M-2									1,00	0,78
M-3										1,00

Tableau 6.5 Accord interobservateur pour la catégorie : Granulaire

Granulaire										
	H-1	**H-2**	**H-3**	**H-4**	**H-5**	**H-6**	**VT-i**	**M-1**	**M-2**	**M-3**
H-1	1,00	0,01	0,03	0,24	0,01	0,00	0,02	0,03	0,02	0,03
H-2		1,00	0,27	-0,01	0,25	0,33	0,23	0,24	0,20	0,31
H-3			1,00	0,02	0,26	0,17	0,19	0,26	0,26	0,19
H-4				1,00	0,03	0,00	-0,01	0,01	0,01	0,00
H-5					1,00	0,29	0,00	0,00	0,10	0,06
H-6						1,00	0,14	0,23	0,29	0,10
VT-i							1,00	0,62	0,48	0,37
M-1								1,00	0,59	0,38
M-2									1,00	0,39
M-3										1,00

Tableau 6.6 Accord interobservateur pour la catégorie : Lisse

Lisse										
	H-1	**H-2**	**H-3**	**H-4**	**H-5**	**H-6**	**VT-i**	**M-1**	**M-2**	**M-3**
H-1	1,00	0,32	0,17	0,32	0,11	0,14	0,26	0,66	0,80	0,00
H-2		1,00	0,65	0,22	0,47	0,27	0,19	0,39	0,26	0,00
H-3			1,00	0,12	0,45	0,34	0,26	0,21	0,33	0,00
H-4				1,00	0,20	0,19	0,42	0,39	0,26	0,00
H-5					1,00	0,54	0,30	0,14	0,09	0,00
H-6						1,00	0,25	0,07	0,21	0,00
VT-i							1,00	0,32	0,48	0,00
M-1								1,00	0,49	0,00
M-2									1,00	0,00
M-3										1,00

Tableau 6.7 Accord interobservateur pour la catégorie : Inconnu

Inconnu										
	H-1	**H-2**	**H-3**	**H-4**	**H-5**	**H-6**	**VT-i**	**M-1**	**M-2**	**M-3**
H-1	1,00	-	-	-	0,00	-	0,00	0,00	0,00	0,00
H-2		1,00	-	-	0,00	-	0,00	0,00	0,00	0,00
H-3			1,00	-	0,00	-	0,00	0,00	0,00	0,00
H-4				1,00	0,00	-	0,00	0,00	0,00	0,00
H-5					1,00	0,00	0,13	-0,01	-0,01	-0,02
H-6						1,00	0,00	0,00	0,00	0,00
VT-i							1,00	-0,04	-0,02	-0,05
M-1								1,00	-0,01	-0,02
M-2									1,00	-0,02
M-3										1,00

Le Tableau 6.8 présente des moyennes de ces mêmes valeurs, afin de nous aider à évaluer les résultats, et une interprétation des coefficients Kappa est proposée dans le Tableau 6.9. En observant les moyennes (Tableau 6.8), nous constatons que la catégorie circulaire semble plus facile à identifier, puisque les moyennes des coefficients Kappa atteignent des accords plus élevés: 0,85 pour les observateurs humains, 0,71 pour les observateurs machines et 0,75 pour tous les observateurs confondus : dans les trois cas, un accord élevé. En observant le détail de ces métriques (Tableau 6.4), nous ne remarquons qu'une légère variabilité : de façon générale, les accords entre les observateurs sont proches.

Nous constatons aussi que les moyennes des coefficients Kappa des catégories parallèle et hachure tendent vers des accords modérés, pour tous les types d'observateurs. Cependant, en observant le détail de ces métriques (Tableau 6.1 et Tableau 6.3), nous remarquons une plus grande fluctuation dans les accords entre les observateurs : par exemple, la catégorie hachure varie de 0,19 (accord très faible) à 0,77 (accord élevé). Ce qui nous laisse croire que l'interprétation de la définition de ces catégories varie parmi les observateurs.

Tableau 6.8 Moyenne des coefficients Kappa

Catégories	**Observateurs**		
	Tous (10)	**Humains (6)**	**Machines (3)**
Parallèle	0,42	0,46	0,39
Arche	0,50	0,76	0,45
Hachure	0,50	0,62	0,43
Circulaire	0,75	0,85	0,71
Granulaire	0,23	0,14	0,24
Lisse	0,22	0,33	0,20
Inconnu	-0,01	0,00	-0,01
Moyenne	0,37	0,45	0,34

Tableau 6.9 Interprétation des coefficients Kappa

Kappa	Interprétation
< 0	Désaccord
0,00	Accord absent (hasard)
0,01 – 0,20	Accord très faible
0,21 – 0,40	Accord faible
0,41 – 0,60	Accord modéré
0,61 – 0,80	Accord élevé
0,81 – 0,99	Accord presque parfait

Quant aux moyennes des coefficients Kappa pour la catégorie arche, nous remarquons qu'elles sont plus élevées pour les observateurs humains, ce qui nous indique qu'elles pourraient être plus difficiles à distinguer par les machines. Nous présumons que les arches pourraient comporter un plus grand nombre d'erreurs dans la vérité de terrain initiale, ce qui pourrait expliquer la performance inférieure des machines. Tout comme pour la catégorie circulaire, nous ne remarquons qu'une légère variabilité dans le détail de ces métriques (Tableau 6.2), ce qui suggère que l'interprétation de la définition de cette catégorie varie moins parmi les observateurs.

Pour la catégorie granulaire, nous observons le contraire : les observateurs machines obtiennent un accord plus élevé que les observateurs humains. Cela suggère que la définition de cette catégorie pourrait être un peu trop floue pour être interprétée de la même façon par différents annotateurs humains. La catégorie lisse présente une moyenne de coefficients Kappa plus élevée pour les observateurs humains, mais cette catégorie est aussi la plus rare, et devrait compter moins d'échantillons. Nous ne tirons pas de conclusions significatives de cette observation.

Plusieurs observateurs n'ont pas identifié d'échantillons comme étant inconnus, et cette métrique ne peut pas être calculée lorsque les deux observateurs de la paire n'ont pas annoté d'échantillons de cette catégorie. Dans les cas où cette métrique est calculée, on note une

absence d'accord (autrement dit, un accord dû au hasard) et même parfois un désaccord. Sauf pour l'observateur humain H-5, qui démontre un accord très faible avec la vérité de terrain initiale.

Tableau 6.10 Nombre d'étiquettes par observateur

Nombre d'étiquettes par observateur										
Nb	**H-1**	**H-2**	**H-3**	**H-4**	**H-5**	**H-6**	**VT-i**	**M-1**	**M-2**	**M-3**
1	14	49	48	17	70	47	49	41	44	32
2	73	48	51	59	29	52	37	46	46	54
3	13	3	1	24	1	1	11	13	10	14
4	0	0	0	0	0	0	2	0	0	0
5	0	0	0	0	0	0	1	0	0	0

Le Tableau 6.10 présente le nombre d'étiquettes identifiées par chaque annotateur. Nous remarquons que seulement deux observateurs humains ont annoté plusieurs échantillons avec trois étiquettes (H-1 et H-4), ce qui nous laisse supposer que les observateurs humains ont tendance à identifier un nombre moins élevé d'étiquettes. De plus, les observateurs machines semblent davantage alignés avec la distribution de la vérité de terrain initiale. Ce qui n'est pas surprenant, vu qu'ils ont été entraînés avec cet ensemble. Finalement, la vérité de terrain initiale inclut trois échantillons avec quatre ou cinq étiquettes, alors qu'aucun autre observateur n'a identifié d'échantillons avec plus de trois étiquettes.

Quelques échantillons étaient accompagnés d'une note dans le fichier d'annotations .csv, provenant du champ de texte libre optionnel de l'outil informatique. En consultant ces informations, nous avons remarqué que plusieurs commentaires mentionnent la provenance de certaines marques, notamment pour la catégorie parallèle. Dans certains cas, les annotateurs précisent que les marques observées semblent provenir de la fabrication de la douille, plutôt que de son utilisation. Ces marques auraient ainsi été causées par des outils d'usinage, et non par les mécanismes d'une arme à feu. Les mêmes échantillons sont souvent mentionnés, indiquant un certain consensus parmi les annotateurs. Des travaux futurs

pourraient évaluer ces commentaires plus en détail, afin d'implémenter une opération additionnelle permettant de faire la distinction entre des marques parallèles provenant de la fabrication de la douille, et des marques parallèles provenant de son utilisation.

Avant de passer à la partie suivante, il convient de souligner que cette étude présente certaines menaces à la validité. Tout d'abord, le fait que tous les annotateurs proviennent de la même entreprise peut amener l'accord observé à refléter des pratiques et conventions internes, plutôt qu'une représentativité plus large de la communauté d'experts. Cette homogénéité des profils risque également de surestimer la robustesse des annotations, en réduisant la diversité des jugements des experts. Par ailleurs, sachant que leurs annotations seraient vérifiées, les participants ont pu adopter une vigilance particulière (effet Hawthorne), ce qui aurait pu influencer leurs jugements. Ils ont, par exemple, pu relever des détails qu'ils n'auraient pas signalés dans des conditions de travail ordinaires, faussant ainsi la mesure de l'accord. Enfin, l'annotation a été réalisée dans un environnement de test, avec un outil offrant moins de possibilités de visualisation que celui utilisé en pratique. Cette différence de conditions a pu modifier la manière dont les experts ont analysé les données. Ces limites invitent donc à interpréter les résultats avec prudence et à les confirmer dans des contextes plus diversifiés et avec des outils plus proches de ceux employés en situation réelle.

6.4.2 Ensemble de vérité de terrain

Cette partie présente les résultats des travaux ayant trait à la création de l'ensemble de vérité de terrain vérifié. Le Tableau 6.11 présente les accords interannotateurs entre la vérité de terrain initiale (VT-i) et la vérité de terrain vérifiée (VT-v), calculés avec le coefficient Kappa, et le Tableau 6.12 présente le nombre d'étiquettes identifiées par chaque annotateur (VT-i et VT-v). Comme attendu, nous ne remarquons qu'un accord faible en moyenne (0,31), entre la vérité de terrain initiale et la vérité de terrain vérifiée. Nous observons que les catégories circulaire et arche atteignent des niveaux d'accord plus élevés que les autres catégories (0,74 et 0,59, respectivement), suggérant encore une fois que ces marques sont

plus faciles à identifier. Les catégories granulaire, lisse et inconnu demeurent des catégories démontrant des accords très faible à absent.

Tableau 6.11 Accord interannotateur entre VT-i et VT-v

Catégories	**Coefficient Kappa**	**Interprétation**
Parallèle	0,47	Accord modéré
Arche	0,59	Accord modéré
Hachure	0,22	Accord faible
Circulaire	0,74	Accord élevé
Granulaire	0,12	Accord très faible
Lisse	0,05	Accord très faible
Inconnu	0,00	Accord absent (hasard)
Moyenne	0,31	Accord faible

Tableau 6.12 Nombre d'étiquettes identifiées par VT-i et VT-v

Nb étiquettes	**VT-i**	**VT-v**
1	537	808
2	407	192
3	55	0
4	1	0
5	0	0

En ce qui concerne le nombre d'étiquettes, nous remarquons la présence d'échantillons annotés avec trois ou quatre étiquettes dans l'ensemble VT-i, alors que l'ensemble VT-v ne comporte aucun échantillon annoté avec plus de deux étiquettes. Finalement, nous observons que la majorité des échantillons sont annotés avec une seule étiquette dans l'ensemble VT-v, alors que dans l'ensemble initial (VT-i), seulement la moitié des échantillons le sont.

Avant de poursuivre avec les expérimentations, nous avons effectué une analyse exploratoire des 1000 échantillons de la vérité de terrain vérifiée (VT-v). Comme attendu, nous avons toujours un ensemble de données déséquilibré.

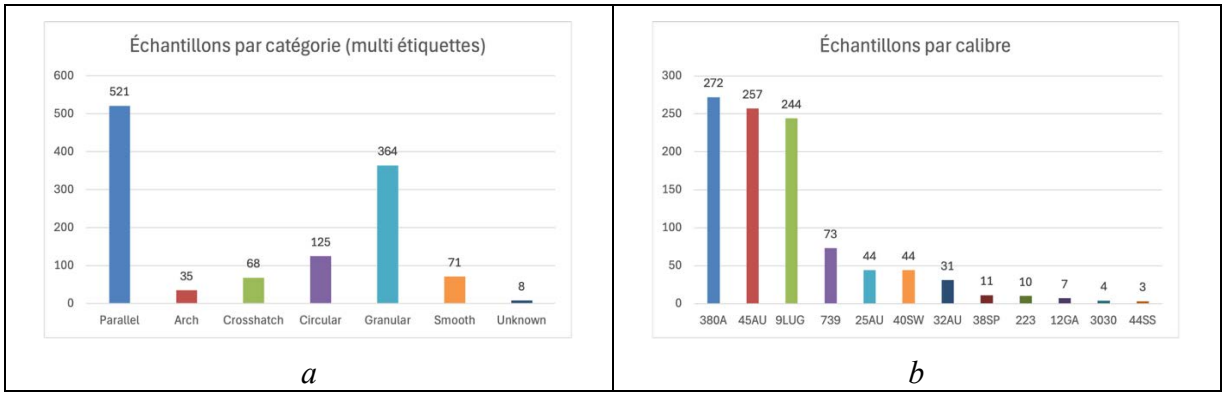


Figure 6.3 BrassTRAX HW4, VT-v : nombre d'échantillons par catégorie (a), nombre d'échantillons par calibre (b)

En observant le graphique de la Figure 6.3 (a), nous remarquons que les catégories parallèle et granulaire sont beaucoup plus présentes que les autres catégories, avec 521 et 364 échantillons affichant ces types de marques. Le graphique de la Figure 6.3 (b) est ajouté à titre informatif, afin de présenter la distribution des calibres parmi les échantillons. Nous pouvons y observer que les calibres 380A, 45AU et 9LUG sont les plus représentés avec 272, 257 et 244 échantillons, respectivement. Le graphique de la Figure 6.4 présente le nombre d'échantillons par combinaison de catégories.

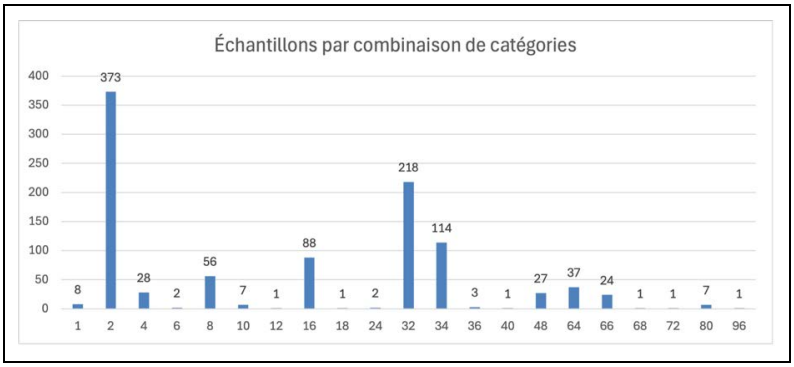


Figure 6.4 BrassTRAX HW4, VT-v : nombre d'échantillons par combinaison de catégories

6.4.3 Classification Multiétiquette

Cette section présente les résultats de la classification multiétiquette. La Figure 6.5 montre les courbes d'apprentissage pour la performance et l'optimisation obtenues lors de l'entraînement du modèle utilisant l'ensemble de vérité de terrain vérifié (VT-v), sans augmentation. Afin de faciliter la comparaison des résultats, nous avons aussi affiché les courbes pour ces mêmes expérimentations utilisant l'ensemble de vérité de terrain initial (VT-i), qui ont été présentées dans le premier objectif.

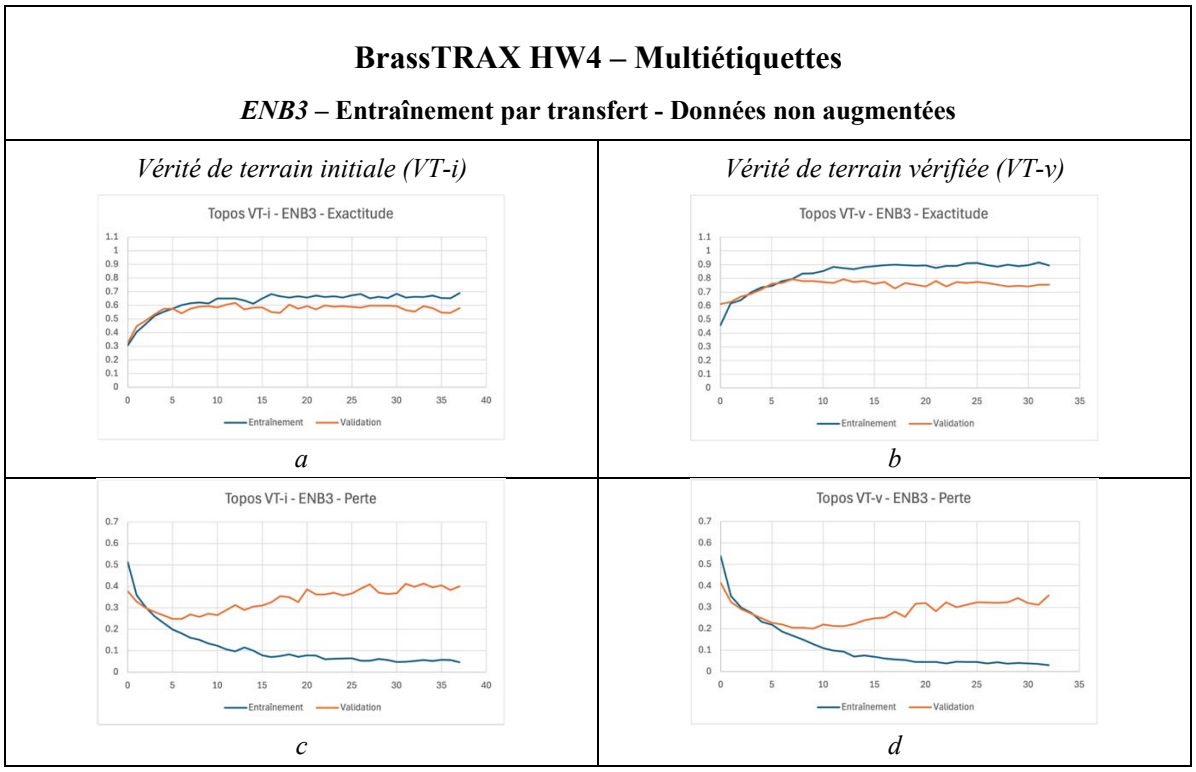


Figure 6.5 Courbes de performance (haut) et courbes d'optimisation (bas) des modèles multiétiquettes, entraînés à partir des données non augmentées : VT-i (a, c) et VT-v (b, d)

Nous remarquons que la performance du modèle VT-v surpasse celle du modèle VT-i : autour de la quinzième époch, nous pouvons observer que la courbe d'entraînement du modèle VT-v (Figure 6.5 b) atteint un plateau dans les environs de 90% d'exactitude, alors que cette même courbe n'atteint pas 70% pour le modèle VT-i (Figure 6.5 a). Cependant, l'écart entre la courbe d'entraînement et celle de validation semble plus grand pour le modèle

VT-v, suggérant une plus grande difficulté à généraliser. Cela pourrait s'expliquer par le nombre inférieur d'échantillons. À titre de rappel, l'ensemble VT-i contient 2311 échantillons, divisés en trois ensembles : 1669 échantillons pour l'ensemble d'entraînement, 347 échantillons pour l'ensemble de validation et 295 échantillons pour l'ensemble de test (voir schéma de la Figure 4.3). Du côté de l'optimisation, nous remarquons que la courbe de validation du modèle VT-v (Figure 6.5 d) est inférieure à celle du modèle VT-i (Figure 6.5 c), réduisant l'écart entre les courbes d'entraînement et de validation et suggérant moins d'erreurs lors de la généralisation pour le modèle VT-v.

Les tableaux qui suivent présentent les résultats obtenus lors de l'évaluation des modèles avec les ensembles de test. Le Tableau 6.13 présente la métrique d'exactitude et la perte, obtenues avec les deux modèles. Nous pouvons observer que l'exactitude augmente avec le modèle VT-v: de 52,88% (VT-i), à 80,31% (VT-v); alors que la perte diminue : de 0,36 (VT-i), à 0,19 (VT-v). Les métriques de précision, de rappel et la mesure-F1 sont présentées dans le Tableau 6.14 pour le modèle VT-i et dans le Tableau 6.15 pour le modèle VT-v.

Tableau 6.13 Métriques d'évaluation des modèles multiétiquettes ENB3

	VT-i	VT-v	VT-v (seuil 65%)
Exactitude	52,88%	80,31%	81,10%
Perte	0,36	0,19	0,18

Tableau 6.14 Rapport de classification du modèle multiétiquette ENB3 VT-i, seuil 50%

VT-i	Précision	Rappel	Mesure-F1	Support
Parallèle	80,49%	61,11%	69,47%	108
Arche	28,57%	20,00%	23,53%	10
Hachure	70,34%	73,45%	71,86%	113
Circulaire	68,09%	86,49%	76,19%	37
Granulaire	87,22%	84,41%	85,79%	186
Lisse	0,00%	0,00%	0,00%	3
Moyenne pondérée	78,05%	74,40%	75,79%	457

Tableau 6.15 Rapport de classification du modèle multiétiquette ENB3 VT-v, seuil 50%

VT-v	Précision	Rappel	Mesure-F1	Support
Parallèle	88,41%	91,04%	89,71%	67
Arche	100,00%	75,00%	85,71%	4
Hachure	0,00%	0,00%	0,00%	8
Circulaire	92,31%	80,00%	85,71%	15
Granulaire	75,56%	73,91%	74,73%	46
Lisse	0,00%	0,00%	0,00%	9
Inconnu	0,00%	0,00%	0,00%	1
Moyenne pondérée	74,56%	73,33%	73,84%	150

Tableau 6.16 Rapport de classification du modèle multiétiquette ENB3 VT-v, seuil 65%

VT-v (seuil 65%)	Précision	Rappel	Mesure-F1	Support
Parallèle	80,26%	91,04%	85,31%	67
Arche	100,00%	75,00%	85,71%	4
Hachure	100,00%	25,00%	40,00%	8
Circulaire	92,31%	80,00%	85,71%	15
Granulaire	70,69%	89,13%	78,85%	46
Lisse	50,00%	22,22%	30,77%	9
Inconnu	0,00%	0,00%	0,00%	1
Moyenne pondérée	77,76%	80,67%	77,12%	150

Ces métriques sont mesurées séparément pour chacune des catégories, et différentes moyennes peuvent ensuite être calculées. Pour l'objectif 1, nous avons présenté la moyenne macro, une moyenne non pondérée des métriques de chaque étiquette, qui ne tient pas compte du déséquilibre. Ici, puisque les données n'ont pas été augmentées dans le but de pallier au déséquilibre, nous préférons observer la moyenne pondérée. Pour trois catégories, nous remarquons une amélioration de la mesure-F1 avec le modèle VT-v : parallèle (de 69.47% pour VT-i, à 89.71% pour VT-v), arche (de 23.53% pour VT-i, à 85.71% pour VT-v)

et circulaire (de 76.19% pour VT-i, à 85.71% pour VT-v). Cependant, la mesure-F1 de la catégorie granulaire demeure supérieure avec le modèle VT-i (de 85.79% pour VT-i, à 74.73% pour VT-v). La mesure-F1 de la catégorie hachure demeure aussi supérieure avec le modèle VT-i, mais aucune prédiction de hachure n'a été faite par le modèle VT-v (de 71.86% pour VT-i, à 0% pour VT-v). Les catégories lisse et inconnu quant à elles, n'ont pas obtenu de prédictions et leur mesure-F1 est de 0% partout.

Malgré tout, la moyenne pondérée du modèle VT-i demeure supérieure à celle du modèle VT-v. Il faut toutefois considérer que le modèle VT-i a seulement une catégorie sans prédiction (lisse), et cette catégorie contient 3 échantillons sur 457, alors que le modèle VT-v en a trois (hachure, lisse et inconnu), contenant 8, 9 et 1 échantillons sur 150. Comme ces catégories sans prédictions influencent la moyenne négativement, il est difficile d'analyser la signification de cette baisse.

En observant le détail des résultats de classification du modèle VT-v, nous remarquons que six (6) échantillons n'ont pas du tout de prédictions, c'est-à-dire qu'une valeur de 0 a été attribuée pour toutes les catégories de ces échantillons. Comme ce modèle comporte une fonction d'activation sigmoïde en sortie, combinée à une perte d'entropie croisée binaire, la prédiction prend la forme d'un vecteur de probabilités, dont la longueur correspond au nombre de catégories et dont la somme n'a pas besoin d'être égale à 1. Cela permet d'établir un seuil afin d'évaluer si chaque catégorie est présente sur l'image : une valeur de probabilité égale ou supérieure au seuil indique que la catégorie est présente et une valeur de prédiction de 1 est attribuée pour cette catégorie, alors qu'une valeur de probabilité inférieure au seuil indique que la catégorie est absente et une valeur de 0 est attribuée.

Jusqu'à maintenant nous avons utilisé un seuil de 50%, une valeur couramment utilisée par défaut. En modifiant ce paramètre à 65%, nous avons réussi à éliminer les prédictions manquantes, tout en augmentant légèrement la plupart des métriques, sauf pour la catégorie parallèle, dont la mesure F1 diminue légèrement. Une colonne VT-v (seuil de 65%) contenant l'exactitude, qui augmente à 81,10% et la perte, qui baisse à 0,18 ont été ajoutées

au Tableau 6.13. Le Tableau 6.16 présente les métriques de précision, de rappel et la mesure-F1 pour le modèle VT-v avec un seuil de 65%. On peut y observer que la mesure F1 des catégories hachure et granulaire augmente à 40,00% et 78,85% par rapport au modèle VT-v avec un seuil de 50%, mais demeurent inférieures à celles du modèle VT-i. La mesure F1 de la catégorie lisse augmente à 30,77%, et dépasse la mesure F1 du modèle VT-i. La mesure F1 des catégories arche et circulaire demeure la même avec 85,71% toutes les deux, et la mesure F1 de la catégorie parallèle baisse à 85,31%. La catégorie inconnu (qui ne comporte qu'un seul échantillon) reste à 0,00%. Le graphique de la Figure 6.6 présente une comparaison visuelle des moyennes pondérées de toutes les métriques, et les matrices de confusion binaire pour chacune des catégories pour le modèle VT-i (avec un seuil de 50%) et VT-v (avec un seuil de 65%) sont présentées à la Figure 6.7. Nous avons omis la matrice de confusion binaire pour la catégorie inconnu, parce qu'elle ne contenait pas d'information utile étant donné qu'il n'y a pas eu de prédictions pour cette catégorie.

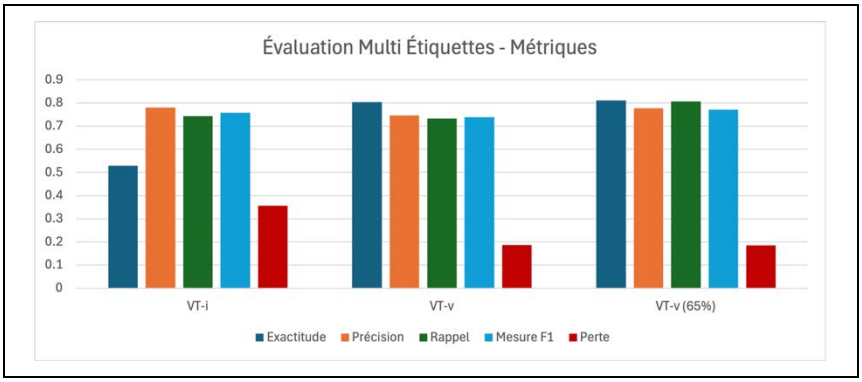


Figure 6.6 Graphique comparatif des métriques d'évaluation et de la perte pour les modèles multiétiquettes VT_i et VT-v

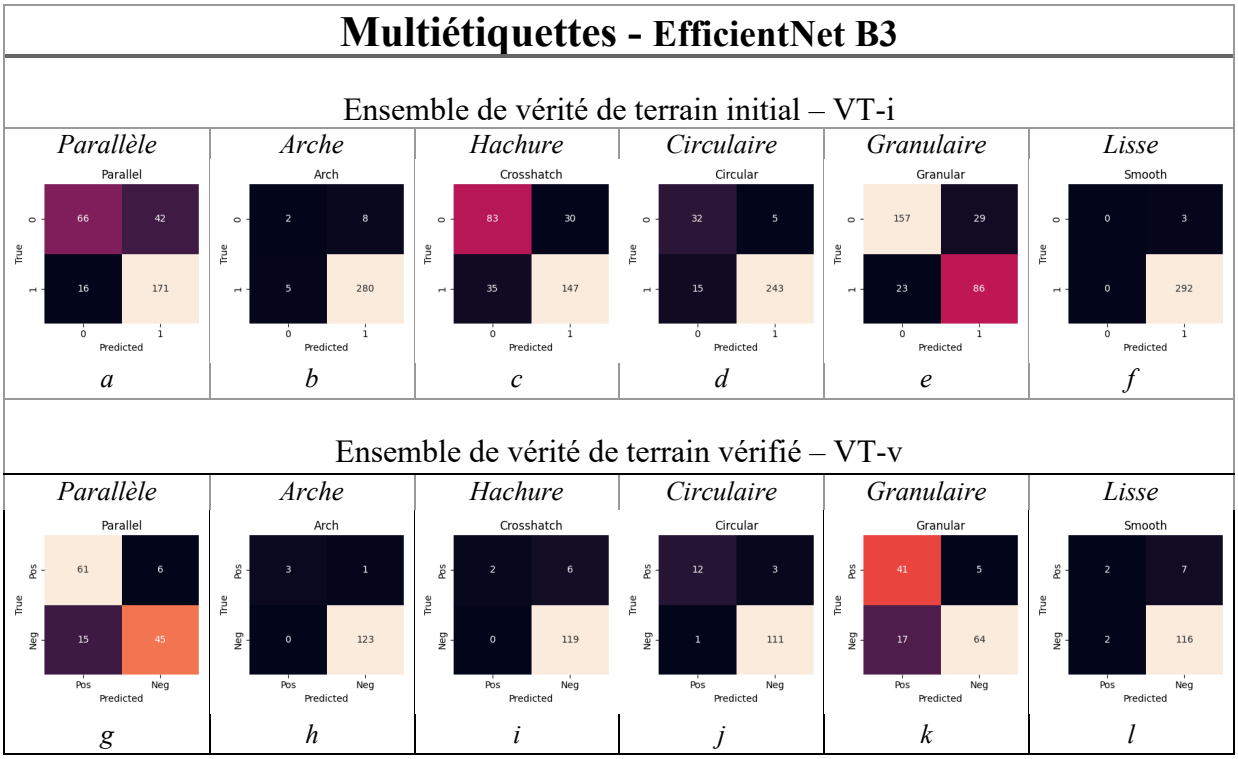


Figure 6.7 Matrices de confusion binaires pour les modèles EfficientNet B3 multiétiquettes : entraînés avec les ensembles VT-i (a-f) et VT-v – seuil 65% (g-l)

En observant les matrices de confusion pour la catégorie parallèle (Figure 6.7 a et g), nous observons 58 erreurs de prédiction pour 295 échantillons pour le modèle VT-i, alors que pour le modèle VT-v, nous observons 21 erreurs pour 127 échantillons. Ces informations soutiennent ce que nous avons observé avec la mesure F1 pour cette catégorie. Nous pouvons aussi observer une baisse dans les erreurs de prédictions pour les catégories arche et circulaire (Figure 6.7 b-h et d-j), les deux autres catégories où nous avons observé une amélioration de la mesure F1 avec le modèle VT-v. Ainsi nous remarquons qu'en général, le modèle VT-v semble mieux adapté que le modèle VT-i pour identifier les échantillons comportant des marques des catégories parallèles, arche et circulaire. Nous remarquons aussi que pour les catégories parallèle, hachure et circulaire, le modèle VT-v comporte davantage d'erreurs de type faux positifs (c'est-à-dire que le modèle a tendance à identifier certaines marques lorsqu'elles ne sont pas présentes); alors que le modèle VT-i semble comporter un plus grand nombre d'erreurs de type faux négatifs (c'est-à-dire que le modèle a tendance à ne

pas identifier certaines marques présentes). Nous observons l'inverse pour la catégorie granulaire.

La Figure 6.8 présente les cartes thermographiques GradCAM++ de quelques exemples pour chacun des deux modèles. En observant ces cartes, nous remarquons que le modèle VT-v (b) semble s'attarder davantage sur la région qui nous intéresse (la marque de la face de la culasse, BF). Nous arrivons à cette conclusion à cause de la forte présence des couleurs rouge, orange et jaune dans cette région. Dans les cartes thermographiques obtenues par le modèle VT-i (a), nous observons beaucoup moins de couleur dans ces régions, voire pas du tout dans certains cas.

Pour chaque catégorie, des exemples d'images 2D venant de vrais positifs, de faux positifs et de faux négatifs sont présentés en annexe (ANNEXE XII, Figure-A XII-1 à Figure-A XII-6). Nous avons omis les exemples appartenant aux vrais négatifs, car nous avons jugé qu'ils n'apportaient pas d'information utile. Par exemple, sur les images de la Figure-A XII-1 (b) et Figure-A XII-2 (c), nous pouvons voir des marques de la catégorie arche, faussement identifiées par le modèle comme étant des marques de la catégorie parallèle. Sur l'image du centre de la Figure-A XII-1 (b), on peut observer un cas où les marques parallèles proviennent de la fabrication de la douille, et non de son utilisation par l'arme à feu. Pour l'instant, les modèles ne sont pas entraînés afin d'en faire la distinction. On peut aussi observer quelques échantillons dans la Figure-A XII-1 (c), sur lesquels des marques parallèles subtiles ne sont pas identifiées par le modèle lorsque des marques des catégories granulaire ou lisse sont présentes. Sur la Figure-A XII-3 (c), on peut observer des exemples où le modèle confond des marques de la catégorie hachure pour des marques parallèles. Les autres figures de cette annexe montrent des exemples pour les trois autres catégories de marques : circulaire, granulaire et lisse.

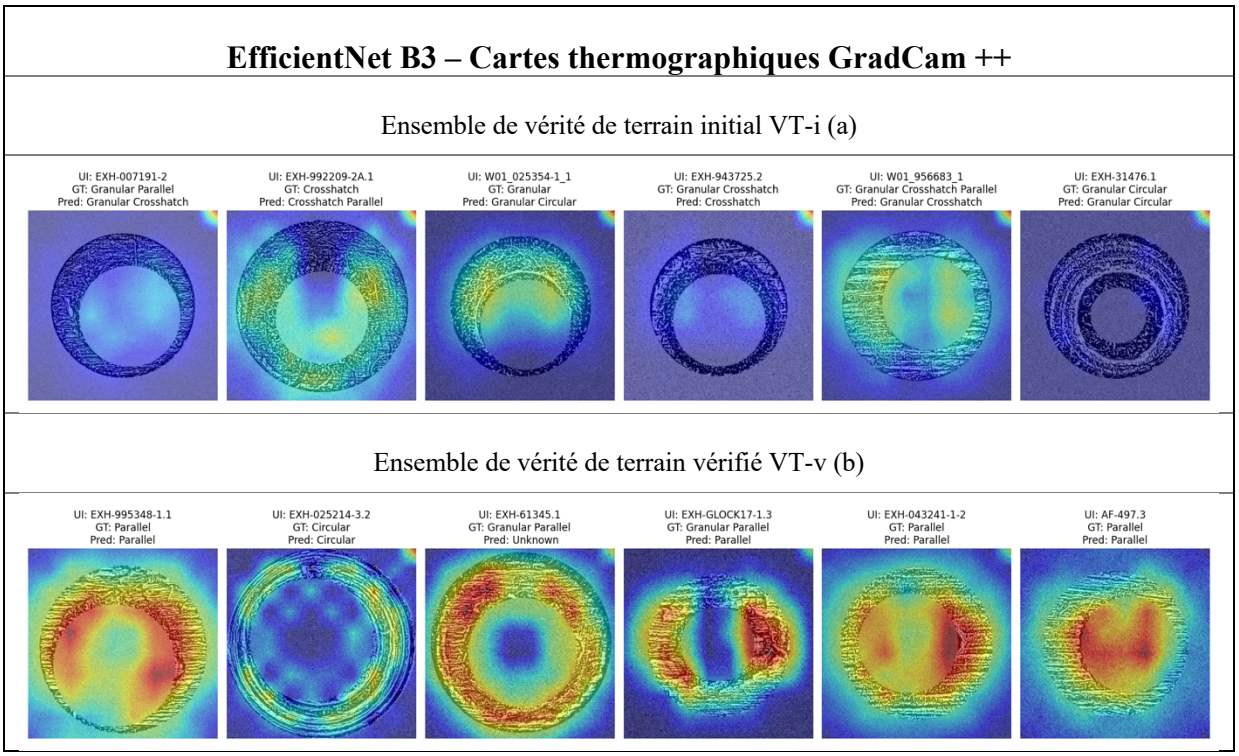


Figure 6.8 Cartes thermographiques pour les modèles EfficientNet B3 multiétiquettes entraînés avec les ensembles VT-i (a) et VT-v (b)

6.5 Conclusion

Dans ce chapitre, nous avons d'abord présenté des métriques interobservateurs, ainsi qu'un nouvel ensemble de vérité de terrain vérifié (VT-v). Afin d'établir les métriques d'accord interobservateurs, nous avons sélectionné 100 échantillons variés. Ensuite, nous avons calculé le coefficient Kappa par paires, entre les annotations de six experts en identification balistique (les observateurs humains), auxquels nous avons ajouté les prédictions des modèles de classification multiétiquette VGG16, ENB3 et ViT B32 (les observateurs machines). Nous avons aussi utilisé la vérité de terrain initiale en tant que 10e observateur. Nous avons constaté que la catégorie circulaire semble plus facile à identifier, avec une moyenne de 0,75 pour les coefficients Kappa, correspondant à un accord élevé. Nous avons observé des accords modérés pour les catégories parallèle, arche et hachure, avec des moyennes des coefficients Kappa de 0,42, 0,50 et 0,50, respectivement; des accords faibles pour les catégories granulaire et lisse, avec des moyennes de 0,23 et 0,22, respectivement; et

un certain désaccord pour la catégorie inconnu, avec une moyenne négative de -0,01. Cependant, nous avons observé une fluctuation entre les accords des observateurs pour les catégories parallèle et hachure, indiquant possiblement que l'interprétation de la définition de ces catégories varie parmi les observateurs. Quant aux moyennes des coefficients Kappa pour la catégorie arche, nous remarquons qu'elles sont plus élevées pour les observateurs humains, ce qui nous indique qu'elles pourraient être plus difficiles à distinguer par les machines. Les catégories arche et circulaire ne montrent qu'une légère variabilité dans les métriques des différents observateurs, ce qui suggère que l'interprétation de la définition de cette catégorie varie moins parmi les observateurs. Pour la catégorie granulaire, nous observons que les observateurs machines obtiennent un accord plus élevé que les observateurs humains. Cela suggère que la définition de cette catégorie pourrait être un peu trop floue pour être interprétée de la même façon par différents annotateurs humains.

Nous avons ensuite recruté deux experts humains pour annoter un ensemble de 1000 échantillons. Ces annotateurs devaient visualiser et discuter des échantillons afin de parvenir à une entente sur les étiquettes, permettant de construire un ensemble de vérité de terrain vérifié. Nous avons observé un accord moyen faible (0,31), entre la vérité de terrain initiale (VT-i) et la vérité de terrain vérifiée (VT-v). Pour le détail, nous avons observé un accord élevé pour la catégorie circulaire (0,74), un accord modéré pour les catégories parallèle (0,47) et arche (0,59), un accord faible pour la catégorie hachure (0,22), un accord très faible pour les catégories granulaire (0,12) et lisse (0,05), et un accord absent pour la catégorie inconnu.

Pour terminer ce chapitre, nous avons repris les expérimentations multiétiquettes avec l'ensemble VT-v, et nous avons comparé les résultats obtenus à ceux du CHAPITRE 4, avec l'ensemble VT-i. Les expérimentations ont été effectuées sans utiliser de techniques d'augmentation sur les données. Nous avons utilisé les échantillons appartenant à la catégorie inconnu pour le modèle VT-v, alors qu'ils avaient été exclus des expérimentations préalables, et nous avons ajouté un seuil à 65%. Nous avons observé une amélioration du modèle, avec une courbe d'entraînement pour la performance atteignant 90,00% d'exactitude pour le

modèle VT-v, alors qu'elle n'atteignait pas 70,00% pour le modèle VT-i. Nous avons aussi observé une augmentation de la métrique d'exactitude : de 52,88% (VT-i) à 80,31% (VT-v) et 81,10% (VT-v seuil de 65%), et une baisse de la perte : de 0,36 (VT-i) à 0,19 (VT-v), et 0,18 (VT-v seuil de 65%). En ce qui concerne les résultats d'évaluation, nous avons observé une variation de la moyenne pondérée pour la mesure F1 avec des valeurs de 75,79% (VT-i), 73,84% (VT-v) et 77,12% (VT-v seuil de 65%). Pour toutes les catégories de ce dernier modèle, la mesure F1 est identique ou augmente, sauf pour la catégorie parallèle où elle diminue. Malgré quelques différences dans l'implémentation (l'ajout de la catégorie inconnue) et dans l'évaluation (le seuil de 65%), nous croyons que le modèle VT-v démontre de meilleures capacités pour la classification des images de douilles.

Cependant, la comparaison avec d'autres études demeure difficile, car les méthodes et les objectifs diffèrent. La majorité des travaux sur les images de douilles de cartouche, dont plusieurs ont été présentés à la section 1.3.2, repose sur des comparaisons par paires de douilles. Par exemple, dans une étude de 2022 (Kara et Karatatar, 2022), les auteurs calculent différents scores de similarité visant à identifier des « matchs » parmi les paires d'échantillons, et classent les dix meilleurs résultats afin d'évaluer la capacité de l'algorithme à retrouver la douille correspondante. Ce type de démarche a pour but d'analyser le comportement d'algorithmes de calcul de similarité. Notre approche s'inscrit dans une logique différente. Plutôt que d'utiliser des comparaisons par paires, le réseau est entraîné sur un vaste ensemble de données dans le but d'apprendre des caractéristiques descriptives permettant de distinguer différentes catégories de marques. Les performances sont ensuite évaluées à l'aide de métriques globales, ce qui reflète la capacité du modèle à généraliser ses prédictions.

Des travaux futurs pourraient inclure des modèles de classification binaire séparés pour chacune des six catégories, ainsi qu'une évaluation des commentaires d'annotations, afin d'implémenter des opérations additionnelles, telles qu'une distinction entre des marques parallèles provenant de la fabrication de la douille, et des marques parallèles provenant de son utilisation.

CONCLUSION

L'objectif principal de cette thèse était de déterminer une méthode permettant d'identifier des marques microscopiques présentes sur des images de douilles de cartouche. Les scènes de crimes impliquant des armes à feu doivent être analysées attentivement en vue de présenter un ensemble de preuves lors d'un procès. Il s'avère utile pour les enquêtes de pouvoir associer une arme à feu avec une scène de crime. Lors du tir d'une cartouche, les mécanismes de l'arme à feu laissent des marques sur la balle et la douille. Ces marques peuvent être des caractéristiques de classe, qui sont communes aux armes à feu d'une même famille; ou des caractéristiques individuelles, qui sont théoriquement uniques et permettent de relier un spécimen à une arme à feu précise. Lorsque des douilles sont retrouvées sur une scène de crime, leurs marques sont analysées dans un laboratoire. Actuellement, avant de comparer les caractéristiques individuelles pour proposer des correspondances, un triage manuel est effectué en considérant les caractéristiques de classe afin d'éviter de calculer inutilement les comparaisons avec des spécimens produits par des familles d'armes différentes. Ce projet a utilisé l'apprentissage machine afin de tenter de détecter automatiquement ces marques.

La première contribution s'intéresse à l'apprentissage supervisé afin de classifier les marques microscopiques présentes sur les douilles de cartouche. La tâche consiste à identifier six catégories de marques (parallèle, arche, hachure, circulaire, granulaire et lisse) et une septième catégorie pour les échantillons avec des marques de type inconnu. Pour cet objectif, nous avons présenté des expérimentations avec des modèles multiclasses, des modèles multiétiquettes, et des modèles binaires. Pour les modèles multiclasses, nous avons obtenu des résultats passables, avec le modèle ENB3, entraîné à partir des données HW4, augmentées et équilibrées : 77,85% pour l'exactitude, 71,53% pour la mesure F1, mais avec une valeur de perte élevée à 1,20. Pour les modèles multiétiquettes, nous avons obtenu des résultats que nous jugeons insuffisants, avec le modèle ENB3, entraîné à partir des données HW4, sans augmentation : 52,88% pour l'exactitude, 54,47% pour la mesure F1, avec une valeur de perte de 0,36. Nous avons constaté que l'augmentation avec une stratégie d'équilibrage des données ne fonctionnait pas bien pour les ensembles multiétiquettes. Les

modèles binaires entraînés à partir des données HW4, sans augmentation, sont ceux qui obtiennent les meilleurs résultats d'évaluation. Les modèles atteignent les valeurs suivantes (pour la mesure F1 et la perte) : parallèle (81% et 0,67); arche (88% et 0,33); hachure (79% et 0,69); circulaire (93% et 0,22); granulaire (80% et 1,02); et lisse (70% et 2,01). Les résultats des classifications nous ont amenés à contester la validité de la vérité de terrain initiale et à mener une brève vérification sur 101 échantillons, évalués par un second expert. Nous avons observé un accord interannotateur variant de modéré à absent, selon la catégorie. Nous avons conclu ce chapitre en suggérant un ré-étiquetage des données de la vérité de terrain. Nous avons aussi offert des pistes pour des travaux futurs utilisant une fonction de perte tenant compte du déséquilibre, telle que la perte focale, afin de tenter d'améliorer les résultats.

La seconde contribution bifurque vers l'apprentissage non supervisé, en évaluant des regroupements d'images de douilles de cartouche. Ces regroupements ont été créés par des algorithmes de clustering traditionnels et un réseau profond de clustering, à partir des caractéristiques descriptives des images extraites par les réseaux entraînés à l'objectif précédent. Au cours de cet objectif, nous avons rencontré plusieurs difficultés. D'abord, les métriques obtenues grâce aux étiquettes des échantillons n'ont pas été particulièrement représentatives des résultats. En partie à cause de la présence d'un nombre élevé d'échantillons étiquetés « inconnu », mais probablement aussi à cause des incohérences de la vérité de terrain qui ont été discutées. Ensuite, le coefficient Silhouette, qui ne dépend pas de la vérité de terrain, est mieux adapté pour les algorithmes formant des clusters sphériques de tailles similaires, et moins bien adaptés lorsque les formes et les tailles des clusters varient. Nous avons aussi rencontré des problèmes au niveau de l'extraction des caractéristiques, que nous avons reliés aux modèles préentraînés sur une vérité de terrain incertaine. Nous nous sommes questionnés sur leur capacité à extraire des caractéristiques pertinentes pour toutes les catégories de marques. Du côté des autoencodeurs et du réseau profond de clustering, nous avons réalisé qu'ils ne parviennent pas à cartographier correctement les informations de texture dans l'espace latent. Nous observons cela à travers la reconstruction des images, qui se concentre principalement sur les formes générales de la région d'intérêt. Des modèles

d'autoencodeurs plus complexes seraient probablement nécessaires pour cette tâche. Nous avons observé que les algorithmes favorisant la création de clusters de formes et de tailles variées semblent mieux adaptés à la représentation de nos données déséquilibrées. Nous avons également apprécié la technique de clustering flou, utilisée par l'algorithme Fuzzy C-Means, qui calcule un degré d'appartenance de chaque point à chaque groupe. Nous avons terminé en suggérant des travaux futurs utilisant une technique de clustering flou avec un algorithme permettant la formation de clusters de tailles variables, tel que le clustering spectral. Finalement, nous avons identifié une autre piste de recherche pour l'utilisation du clustering. Plutôt que d'appliquer le clustering sur les caractéristiques descriptives de l'ensemble des images, il pourrait être utilisé sur chaque image séparément pour regrouper les pixels similaires dans une image afin d'en faire une segmentation des différents éléments, qui pourraient ensuite être classifiés.

Par suite des observations sur les incertitudes de la vérité de terrain, la dernière contribution s'est chargée d'étudier l'accord interobservateur entre des observateurs humains et les méthodes d'apprentissage profond supervisé. Nous avons d'abord présenté des métriques d'accord interobservateur, obtenues sur un ensemble de 100 échantillons variés. Les coefficients Kappa ont été calculés par paires entre les annotations de six experts en identification balistique (les observateurs humains), auquel nous avons ajouté les prédictions des modèles de classification multiétiquettes VGG16, ENB3 et ViT B32 (les observateurs machines). Nous avons aussi utilisé la vérité de terrain initiale en tant que 10e observateur. Nous avons obtenu un accord moyen élevé pour la catégorie circulaire (0,75), un accord moyen modéré pour les catégories parallèle (0,42), arche (0,50) et hachure (0,50), un accord faible pour les catégories granulaire (0,23) et lisse (0,22) et un désaccord pour la catégorie inconnu (-0,01). Nous avons ensuite recruté deux experts humains pour annoter un ensemble de 1000 échantillons. Nous avons observé un accord moyen faible, entre la vérité de terrain initiale (VT-i) et la vérité de terrain vérifiée (VT-v). Pour le détail, nous avons observé un accord élevé pour la catégorie circulaire (0,74), un accord modéré pour les catégories parallèle (0,47) et arche (0,59), un accord faible pour la catégorie hachure (0,22), un accord très faible pour les catégories granulaire (0,12) et lisse (0,05), et un accord absent pour la

catégorie inconnu (0). Pour terminer ce chapitre, nous avons repris les expérimentations multiétiquettes avec l'ensemble VT-v, et nous avons comparé les résultats obtenus à ceux dans la première contribution, avec l'ensemble VT-i. Les expérimentations VT-v ont été effectuées sans utiliser de techniques d'augmentation sur les données, avec les échantillons appartenant à la catégorie inconnu et un seuil de prédiction augmenté à 65%. Nous avons observé une amélioration, avec une courbe d'entraînement pour la performance atteignant 90,00% d'exactitude pour le modèle VT-v, alors qu'elle n'atteignait pas 70,00% pour le modèle VT-i. Nous avons aussi observé une augmentation de la métrique d'exactitude : de 52,88% (VT-i) à 80,31% (VT-v) et 81,10% (VT-v seuil de 65%), et une baisse de la perte : de 0,36 (VT-i) à 0,19 (VT-v), et 0,18 (VT-v seuil de 65%). En ce qui concerne les résultats d'évaluation, nous avons observé une variation de la moyenne pondérée pour la mesure F1 avec des valeurs de 75,79% (VT-i), 73,84% (VT-v) et 77,12% (VT-v seuil de 65%). Somme toute, nous croyons que le modèle VT-v démontre de meilleures capacités pour la classification des images de douilles. Des travaux futurs pourraient inclure des modèles de classification binaire VT-v pour chacune des six catégories, ainsi qu'une évaluation des commentaires d'annotations, afin d'implémenter des opérations additionnelles, telles qu'une distinction entre des marques parallèles provenant de la fabrication de la douille, et des marques parallèles provenant de son utilisation.

En dernière analyse, il ressort de cette thèse que l'apprentissage profond constitue une approche prometteuse pour l'identification automatique de marques descriptives sur des images de douilles de cartouche. En revenant à la problématique de départ, à savoir dans quelle mesure ces méthodes peuvent transformer l'analyse balistique en un processus plus fiable, plus rapide, capable de traiter une quantité croissante de données et moins dépendant de l'interprétation subjective de l'expert, il est possible d'apporter les réponses suivantes aux trois questions de recherche.

À la première question, portant sur la possibilité d'améliorer le processus d'identification balistique grâce à une assistance automatisée de la détection des marques de classe, les résultats suggèrent plusieurs pistes positives. En effet, les modèles de classification

pourraient être intégrés dans des systèmes automatisés d'identification afin d'assister les techniciens lors de la saisie de l'information, notamment en proposant des catégories de marques pertinentes. Ces propositions devraient toutefois être considérées comme de simples suggestions et non comme des réponses définitives, l'expert conservant toujours un rôle décisionnel central avec la possibilité d'accepter, de corriger ou de rejeter la prédiction du modèle. Une telle assistance aurait également pour effet de fournir plus rapidement des informations utiles aux enquêteurs, tout en améliorant l'efficacité des calculs de jumelage grâce à l'élimination en amont des échantillons présentant des types de marques différents.

À la deuxième question, concernant la capacité de l'apprentissage non supervisé à révéler des régularités significatives dans les marques balistiques sans étiquetage préalable, les résultats montrent que cette approche ne s'avère pas concluante pour les marques microscopiques, du moins dans l'état actuel des méthodes. La méthode présente néanmoins un potentiel intéressant pour la reconnaissance de formes. Par exemple, elle pourrait permettre de regrouper les formes de la marque de la face de la culasse, et ainsi fournir une information supplémentaire au système. Ces regroupements pourraient également être exploités comme caractéristiques additionnelles pour alimenter les classificateurs.

Enfin, concernant la troisième question, relative à la variabilité interobservateur et à sa comparaison avec les méthodes automatiques, il apparaît que cette variabilité demeure significative, ce qui souligne l'intérêt de recourir à des approches automatisées pour limiter l'influence de la subjectivité humaine. Cette variabilité est particulièrement marquée pour les catégories parallèle et hachure, tandis que la catégorie circulaire montre un accord élevé. Les modèles entraînés sur la vérité terrain initiale présentent en moyenne des accords légèrement inférieurs à ceux des observateurs humains, à l'exception de la catégorie granulaire. Cependant, il est possible d'améliorer les performances des modèles en s'appuyant sur un ensemble de vérité terrain vérifié.

Il reste toutefois du travail à faire avant de pouvoir inclure ces modèles dans un système automatisé. D'abord, le modèle multiétiquette VT-v, même s'il pourrait détecter les marques

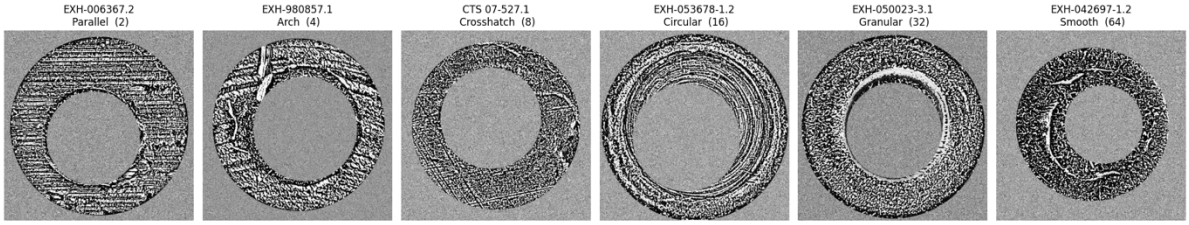
des catégories circulaire et arche, n'est pas tout à fait prêt pour les autres catégories. Les marques parallèles pourraient aussi être détectées, mais les modèles étudiés ne font pas de distinction entre des marques parallèles causées lors de la fabrication de la douille, et des marques parallèles causée par l'utilisation d'une arme à feu. Nous recommandons de poursuivre la recherche avec les classificateurs binaires tout en augmentant l'ensemble de vérité de terrain vérifié, en particulier avec l'ajout d'échantillons appartenant aux catégories plus rares. Les algorithmes de clustering flous pourraient aussi être explorés davantage, mais il sera nécessaire d'améliorer la méthode d'extraction des caractéristiques, par exemple avec des CNN préentraînées avec un ensemble de vérité de terrain vérifié, ou avec des autoencodeurs plus avancés.

ANNEXE I

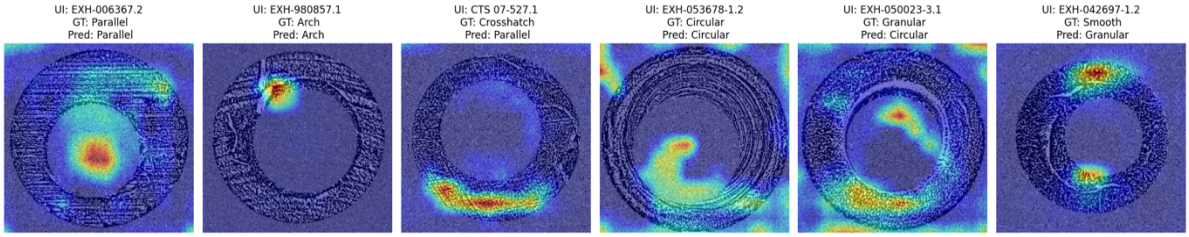
CONTRIBUTION #1 : MULTICLASSE

Cartes de Visualisation

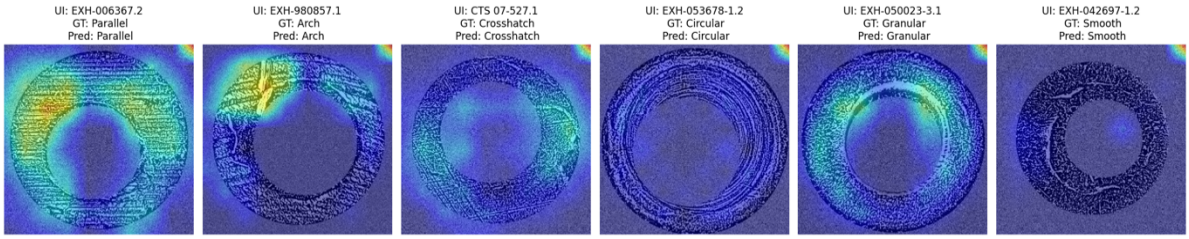
Images originales (a)



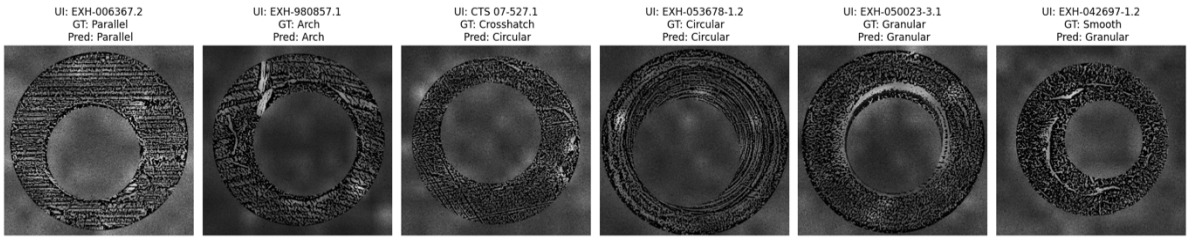
VGG16 – Cartes thermographiques GradCAM ++ (b)



EfficientNet B3 – Cartes thermographiques GradCAM ++ (c)



ViT B16 – Cartes d'attention (d)



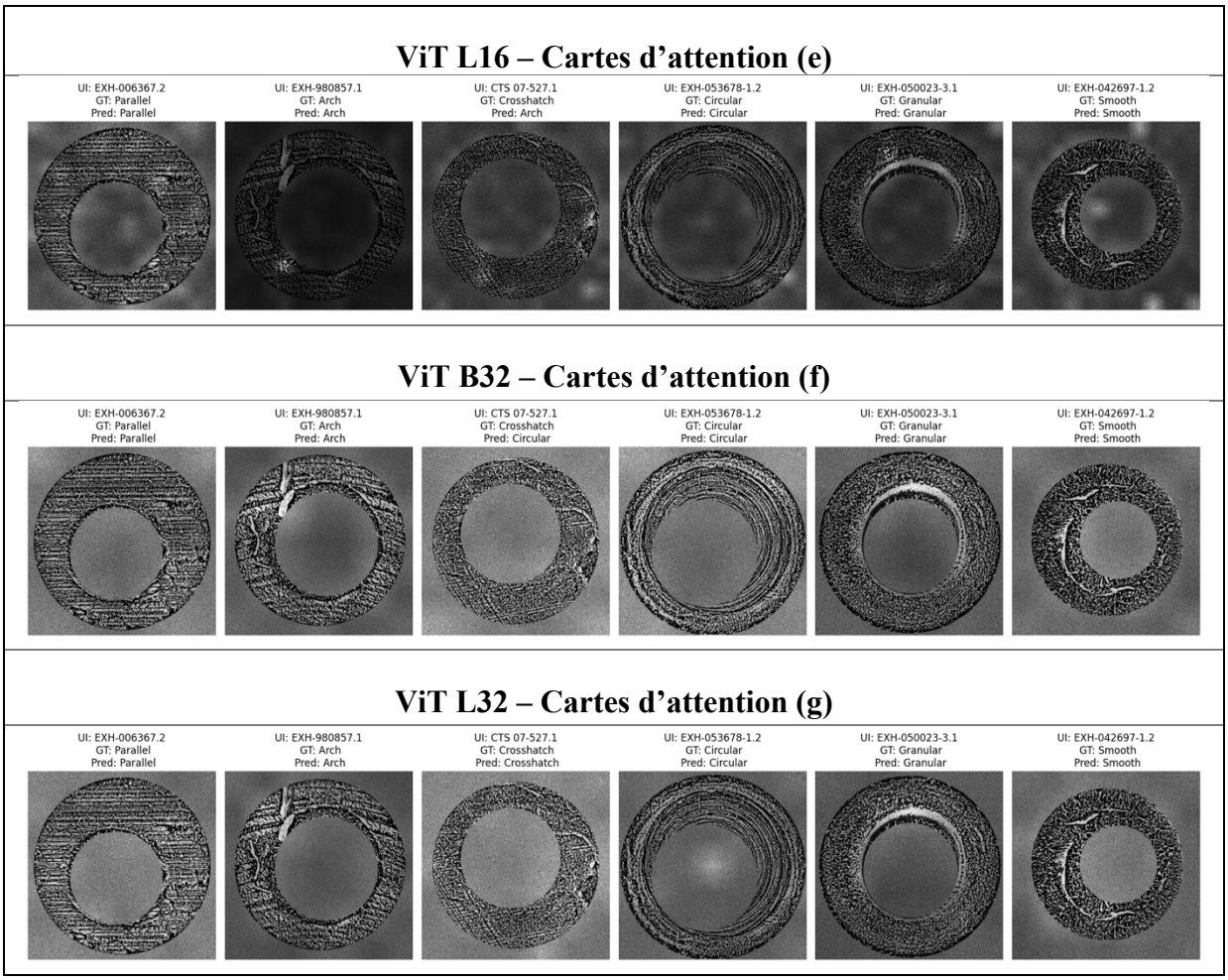


Figure-A I-1 Cartes de visualisation pour les expérimentations multiclasses. Images originales (a) Cartes thermographiques pour VGG16 (b) et EfficienNet B3 (c), et cartes d'attention pour les ViT (d-g)

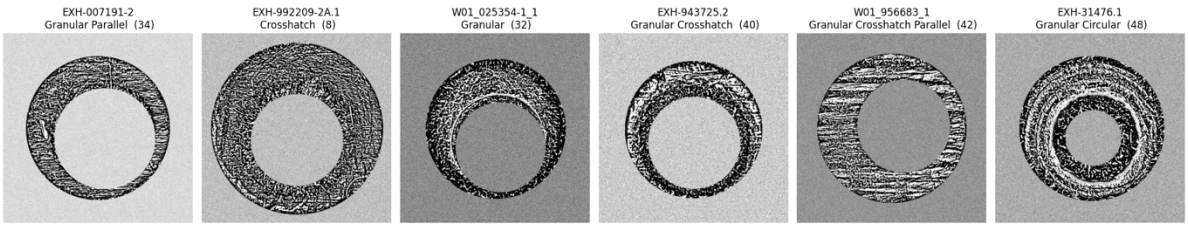
ANNEXE II

CONTRIBUTION #1 : MULTIÉTIQUETTE

Figure-A II-1 Matrices de confusion binaires pour les expérimentations multiétiquettes. VGG16 : avec données augmentées (a), sans augmentation (b). ViT B32 : avec données augmentées (c), sans augmentation (d)

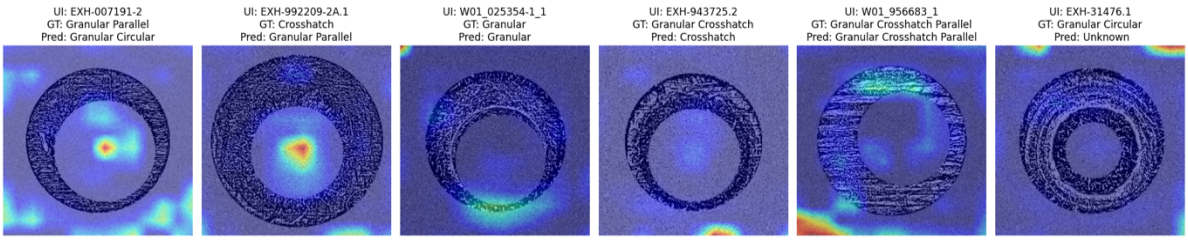
Cartes de Visualisation

Images originales (a)

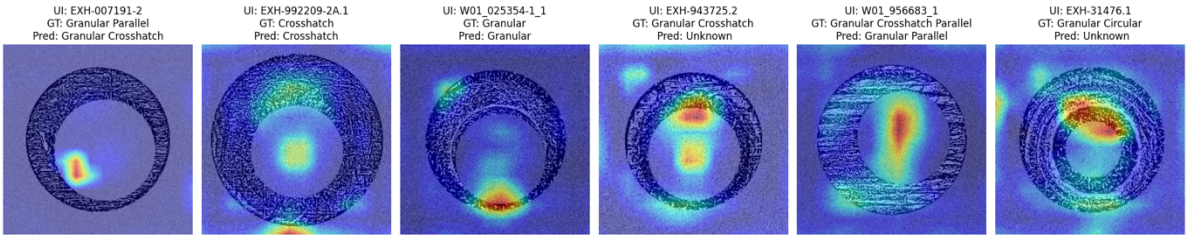


VGG16 – Cartes thermographiques GradCAM ++ (b)

Données augmentées



Données sans augmentation



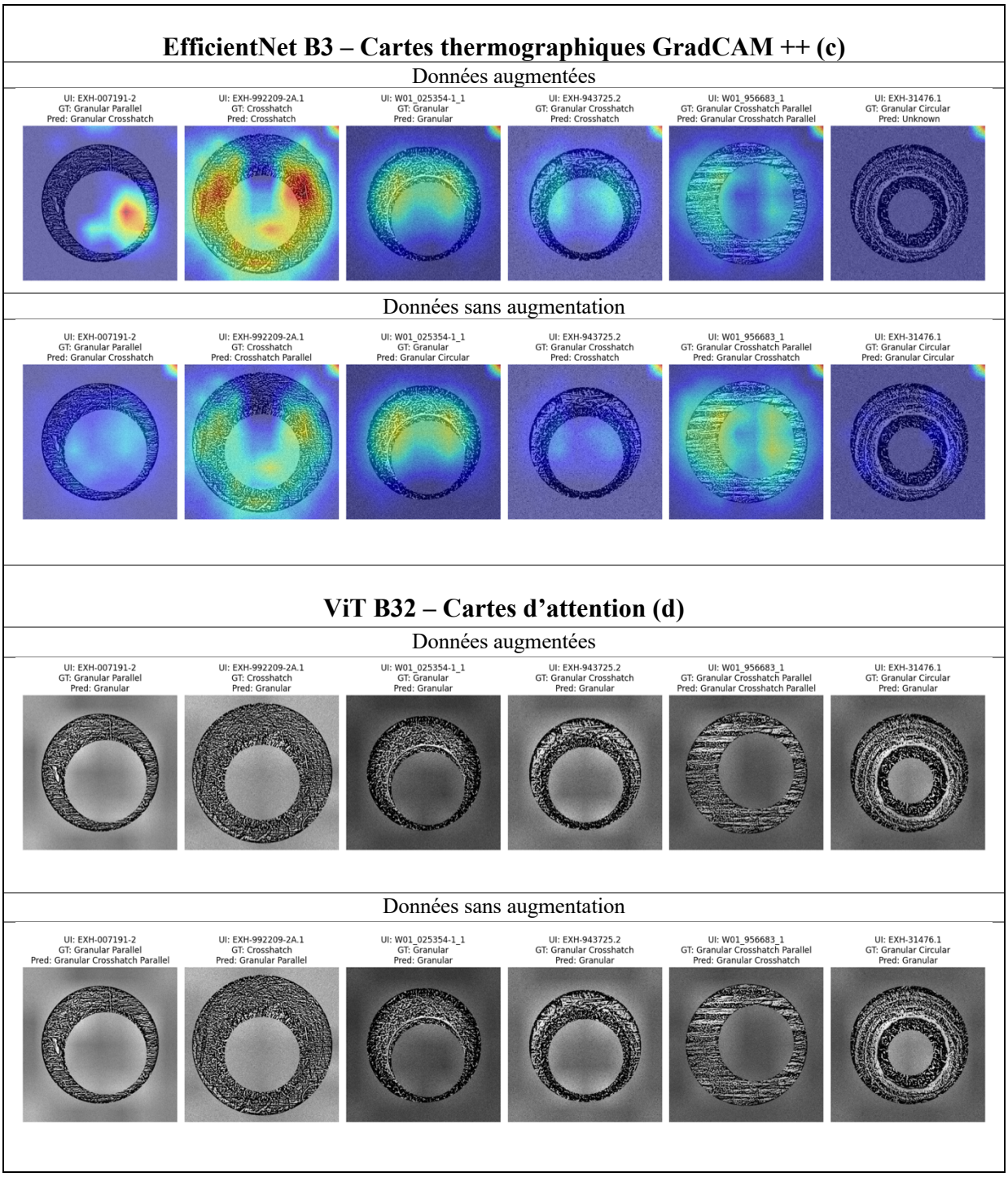


Figure-A II-2 Cartes de visualisation pour les expérimentations multiétiquettes. Images originales (a) Cartes thermographiques pour VGG16 (b) et EfficienNet B3 (c), et cartes d'attention pour les ViT B32 (d)

ANNEXE III

CONTRIBUTION #1 : BINAIRE

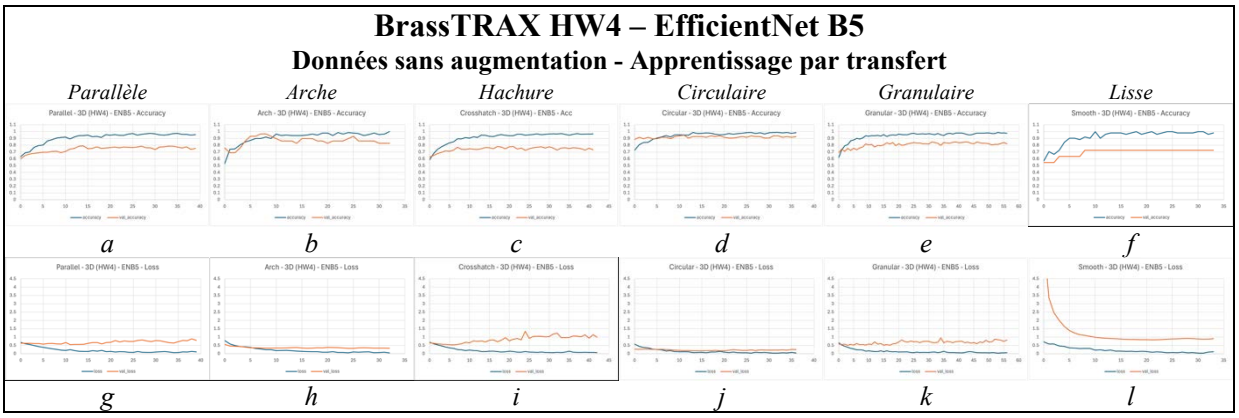


Figure-A III-1 Courbes d'apprentissage du modèle binaire ENB5, données augmentées du BrassTRAX HW4 : courbes de performance (haut) et courbes d'optimisation (bas)

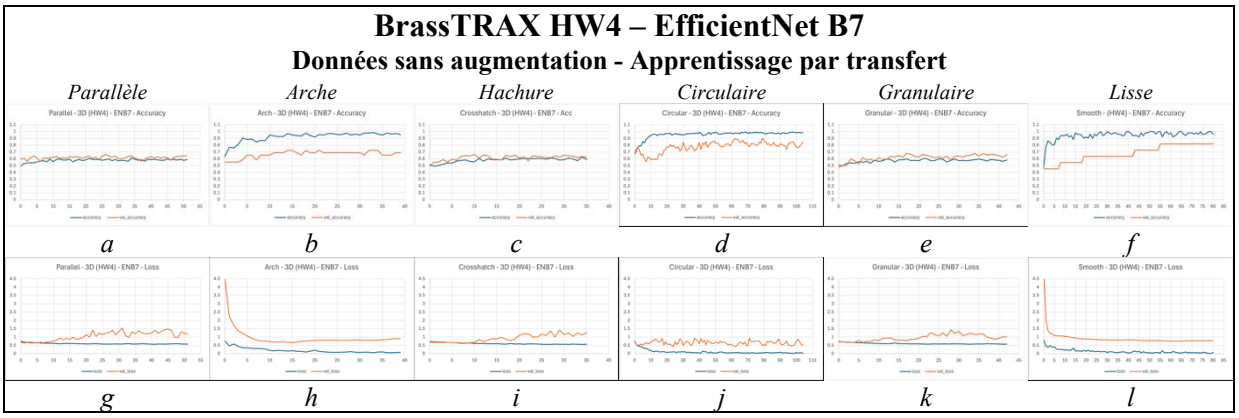


Figure-A III-2 Courbes d'apprentissage du modèle binaire ENB7, données augmentées du BrassTRAX HW4 : courbes de performance (haut) et courbes d'optimisation (bas)

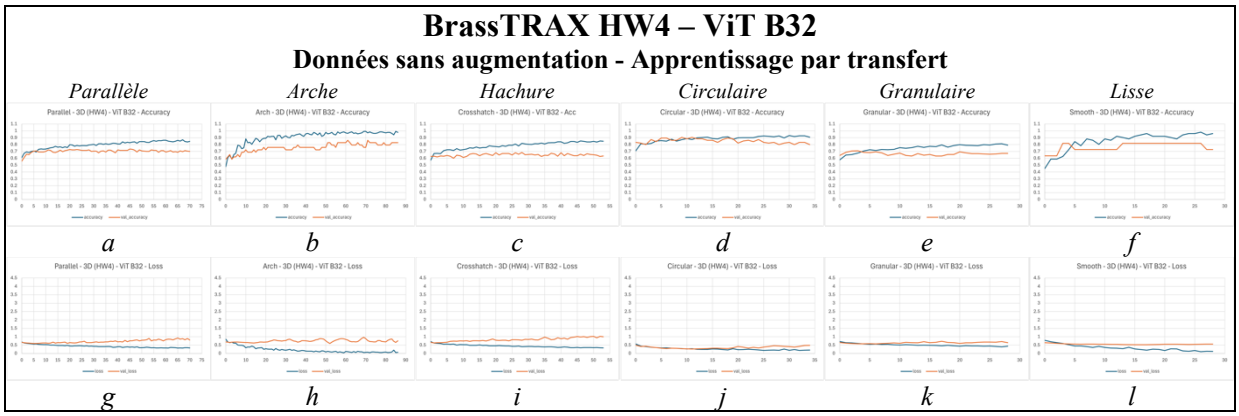


Figure-A III-3 Courbes d'apprentissage du modèle binaire ViT B32, données augmentées du BrassTRAX HW4 : courbes de performance (haut) et courbes d'optimisation (bas)

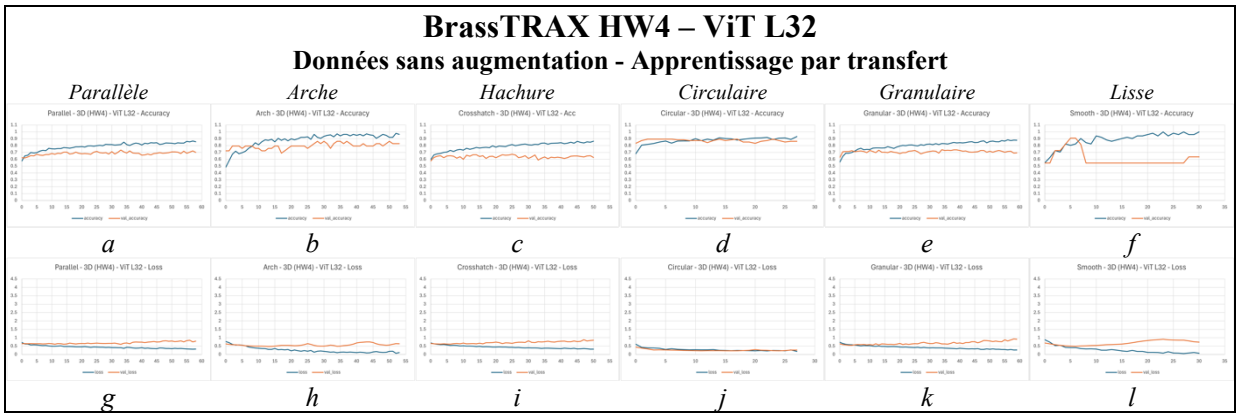


Figure-A III-4 Courbes d'apprentissage du modèle binaire ViT L32, données augmentées du BrassTRAX HW4 : courbes de performance (haut) et courbes d'optimisation (bas)

Tableau-A III-1 Métriques d'évaluation du modèle binaire pour la catégorie « parallèle », données sans augmentation du BrassTRAX HW4

Modèle	EX	PR	Rappel	F1	MCC	Auroc	Perte
ENB5	79,19%	79,43%	79,16%	79,13%	0,59	86,94%	0,59
ENB7	61,09%	61,88%	61,14%	60,50%	0,23	68,47%	1,22
ViT B32	68,33%	68,48%	68,35%	68,27%	0,37	76,87%	0,72
ViT L32	76,02%	76,07%	76,01%	76,00%	0,52	82,61%	0,58

Tableau-A III-2 Métriques d'évaluation du modèle binaire pour la catégorie « arche », données sans augmentation du BrassTRAX HW4

Modèle	EX	PR	Rappel	F1	MCC	Auroc	Perte
ENB5	72,00%	72,12%	72,12%	72,00%	0,44	85,90%	0,45
ENB7	76,00%	78,82%	76,60%	75,65%	0,55	80,13%	0,64
ViT B32	76,00%	79,04%	75,32%	75,00%	0,54	83,33%	0,85
ViT L32	76,00%	76,67%	75,64%	75,65%	0,52	82,69%	0,63

Tableau-A III-3 Métriques d'évaluation du modèle binaire pour la catégorie « hachure », données sans augmentation du BrassTRAX HW4

Modèle	EX	PR	Rappel	F1	MCC	Auroc	Perte
ENB5	75,00%	75,01%	75,00%	75,00%	0,50	83,01%	0,93
ENB7	75,00%	75,93%	75,00%	74,77%	0,51	79,15%	0,57
ViT B32	72,41%	72,47%	72,41%	72,40%	0,45	82,51%	0,62
ViT L32	71,98%	72,36%	71,98%	71,87%	0,44	80,53%	0,58

Tableau-A III-4: Métriques d'évaluation du modèle binaire pour la catégorie « circulaire », données sans augmentation du BrassTRAX HW4

Modèle	EX	PR	Rappel	F1	MCC	Auroc	Perte
ENB5	97,53%	97,53%	97,53%	97,53%	0,95	99,33%	0,11
ENB7	92,59%	92,72%	92,56%	92,58%	0,85	95,91%	0,32
ViT B32	88,89%	90,02%	88,99%	88,83%	0,79	98,60%	0,24
ViT L32	91,36%	91,94%	91,43%	91,34%	0,83	98,54%	0,26

Tableau-A III-5 Métriques d'évaluation du modèle binaire pour la catégorie « granulaire », données sans augmentation du BrassTRAX HW4

Modèle	EX	PR	Rappel	F1	MCC	Auroc	Perte
ENB5	79,73%	79,75%	79,73%	79,73%	0,59	85,59%	0,84
ENB7	62,61%	63,08%	62,61%	62,28%	0,26	66,46%	0,89
ViT B32	64,86%	66,61%	64,86%	63,92%	0,31	72,58%	0,64
ViT L32	69,82%	70,88%	69,82%	69,43%	0,41	79,08%	0,75

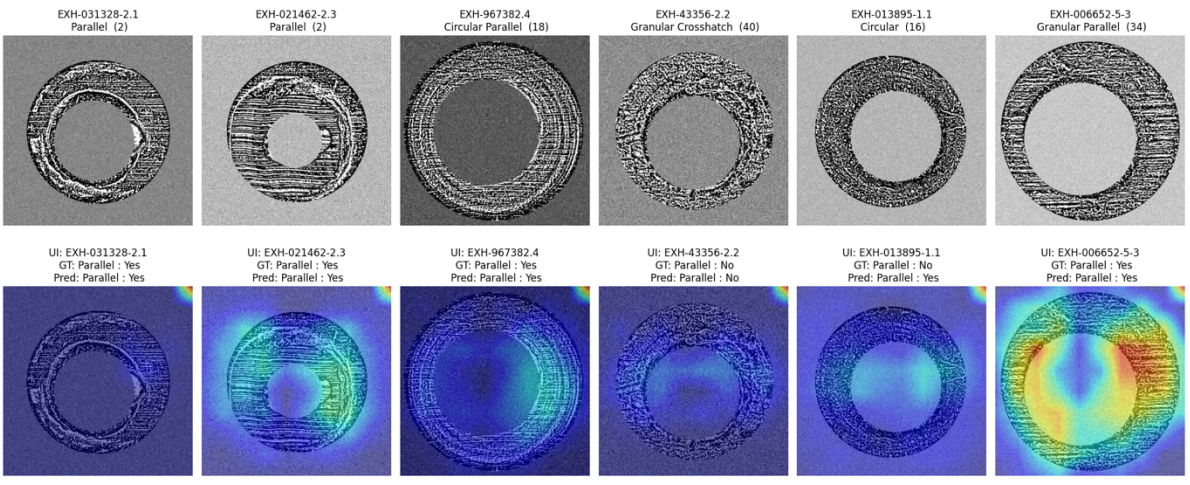
Tableau-A III-6 Métriques d'évaluation du modèle binaire pour la catégorie « lisse », données sans augmentation du BrassTRAX HW4

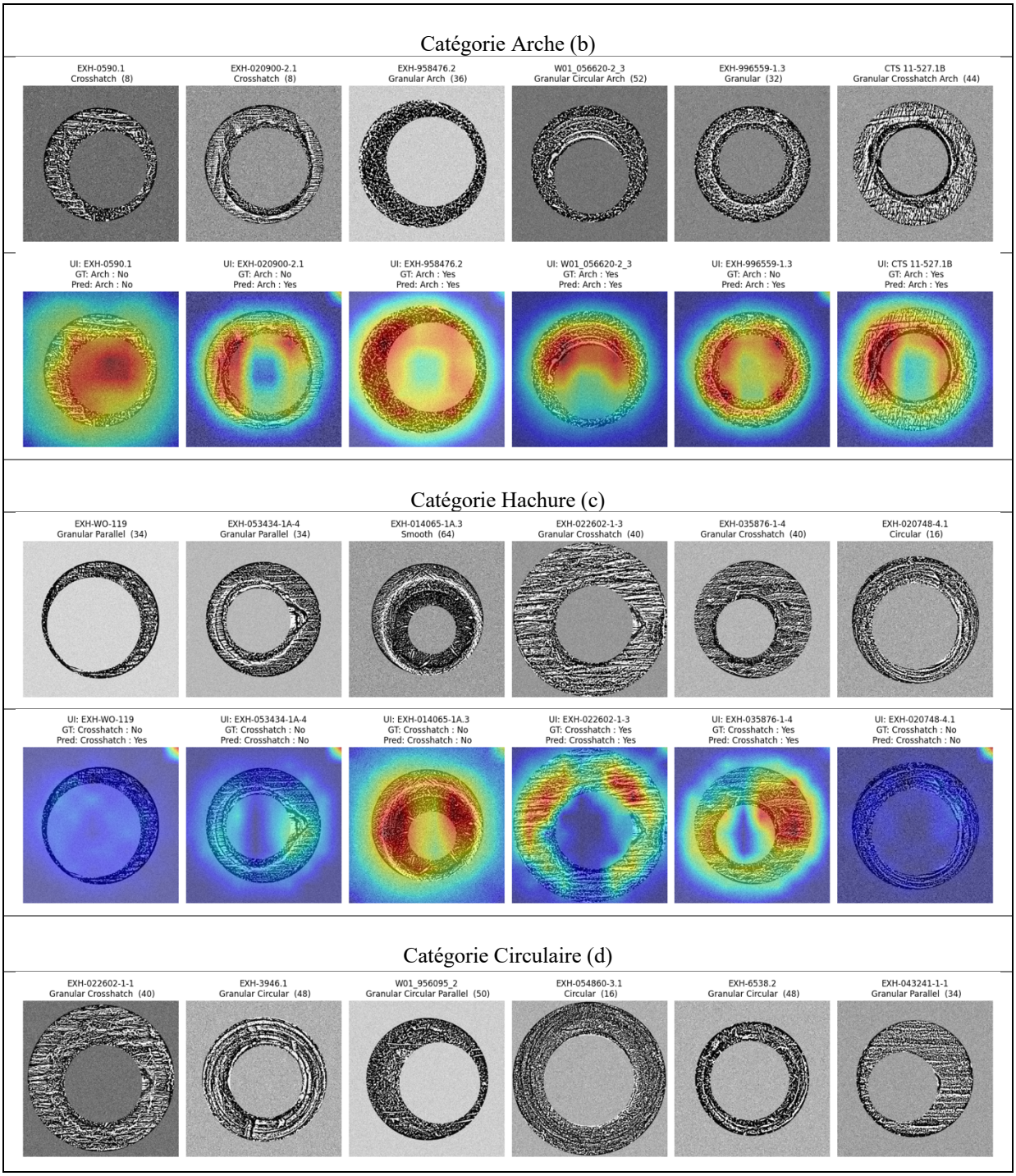
Modèle	EX	PR	Rappel	F1	MCC	Auroc	Perte
ENB5	70,00%	70,83%	70,00%	69,70%	0,41	64,00%	1,47
ENB7	70,83%	70,00%	69,70%	70,00%	0,41	64,00%	1,01
ViT B32	50,00%	50,00%	50,00%	49,49%	0,00	32,00%	0,74
ViT L32	70,00%	81,25%	70,00%	67,03%	0,50	52,00%	0,70

Cartes de Visualisation

EfficientNet B3 – Cartes thermographiques GradCAM ++

Catégorie Parallèle (a)




Catégorie Arche (b)
EXH-0590.1
Crosshatch (8)
EXH-020900-2.1
Crosshatch (8)
EXH-958476.2
Granular Arch (36)
W01_056620-2_3
Granular Circular Arch (52)
EXH-996559-1.3
Granular (32)
CTS 11-527.1B
Granular Crosshatch Arch (44)
UI: EXH-0590.1
GT: Arch : No
Pred: Arch : No
UI: EXH-020900-2.1
GT: Arch : No
Pred: Arch : Yes
UI: EXH-958476.2
GT: Arch : Yes
Pred: Arch : Yes
UI: W01_056620-2_3
GT: Arch : Yes
Pred: Arch : Yes
UI: EXH-996559-1.3
GT: Arch : No
Pred: Arch : Yes
UI: CTS 11-527.1B
GT: Arch : Yes
Pred: Arch : Yes
Catégorie Hachure (c)
EXH-WO-119
Granular Parallel (34)
EXH-053434-1A-4
Granular Parallel (34)
EXH-014065-1A.3
Smooth (64)
EXH-022602-1-3
Granular Crosshatch (40)
EXH-035876-1-4
Granular Crosshatch (40)
EXH-020748-4.1
Circular (16)
UI: EXH-WO-119
GT: Crosshatch : No
Pred: Crosshatch : Yes
UI: EXH-053434-1A-4
GT: Crosshatch : No
Pred: Crosshatch : No
UI: EXH-014065-1A.3
GT: Crosshatch : No
Pred: Crosshatch : No
UI: EXH-022602-1-3
GT: Crosshatch : Yes
Pred: Crosshatch : Yes
UI: EXH-035876-1-4
GT: Crosshatch : Yes
Pred: Crosshatch : Yes
UI: EXH-020748-4.1
GT: Crosshatch : No
Pred: Crosshatch : No
Catégorie Circulaire (d)
EXH-022602-1-1
Granular Crosshatch (40)
EXH-3946.1
Granular Circular (48)
W01_956095_2
Granular Circular Parallel (50)
EXH-054860-3.1
Circular (16)
EXH-6538.2
Granular Circular (48)
EXH-043241-1-1
Granular Parallel (34)

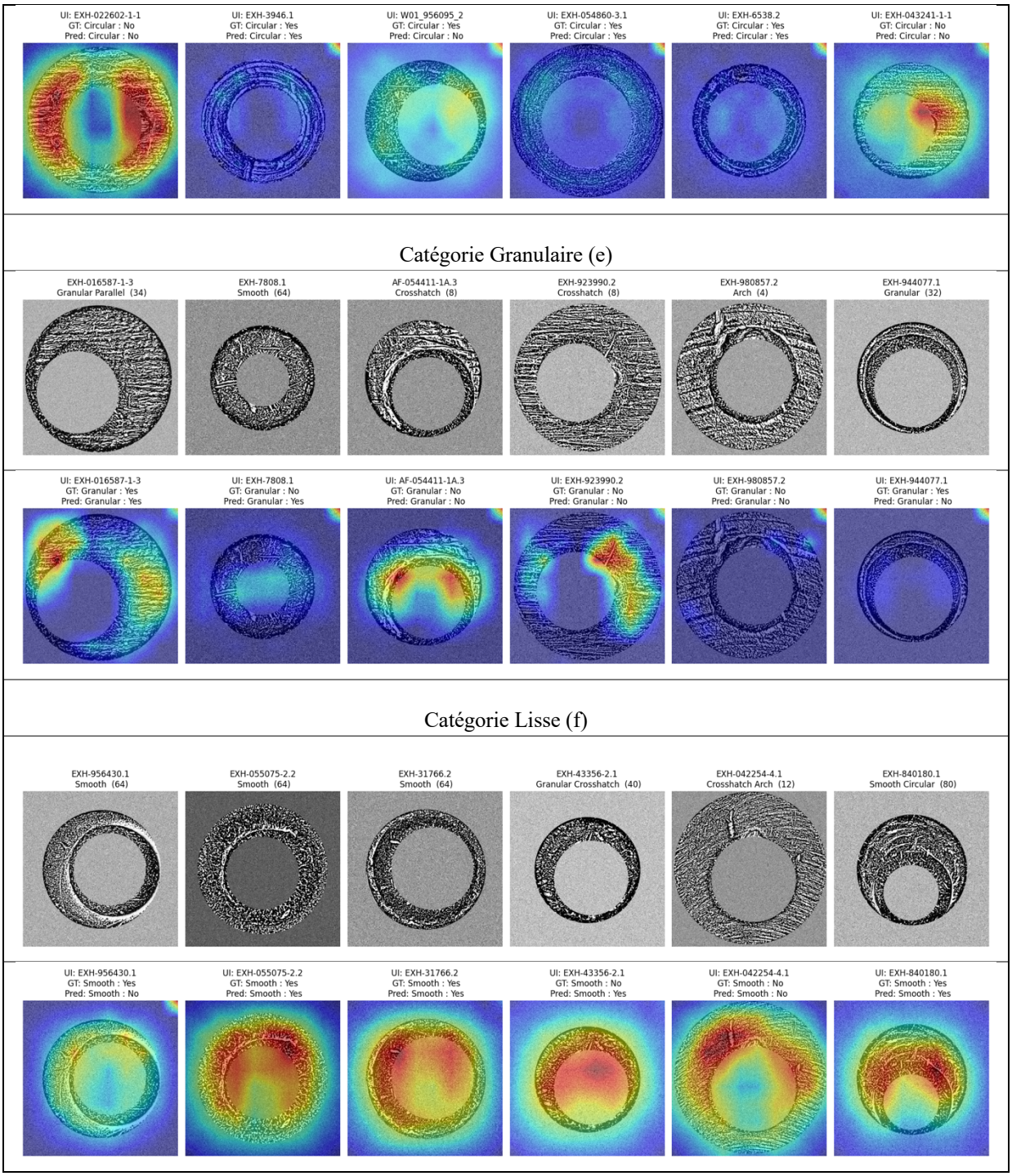


Figure-A III-5 Cartes de visualisation pour les expérimentations binaires : Parallèle (a), Arche (b), Hachure (c), Circulaire (d), Granulaire (e) et Lisse (f). Images originales (haut), cartes thermographiques pour ENB3, données sans augmentation (bas)

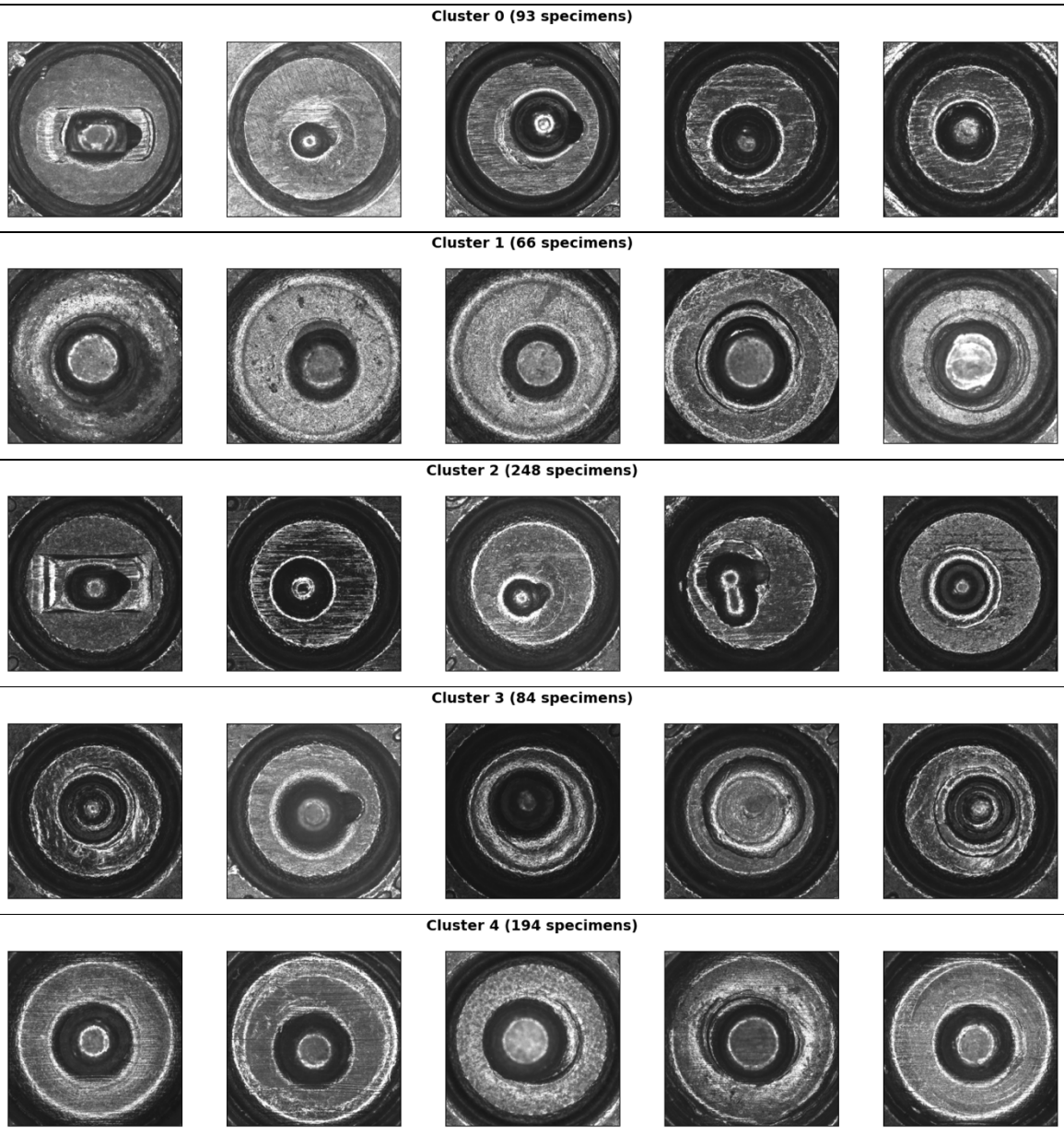
ANNEXE IV

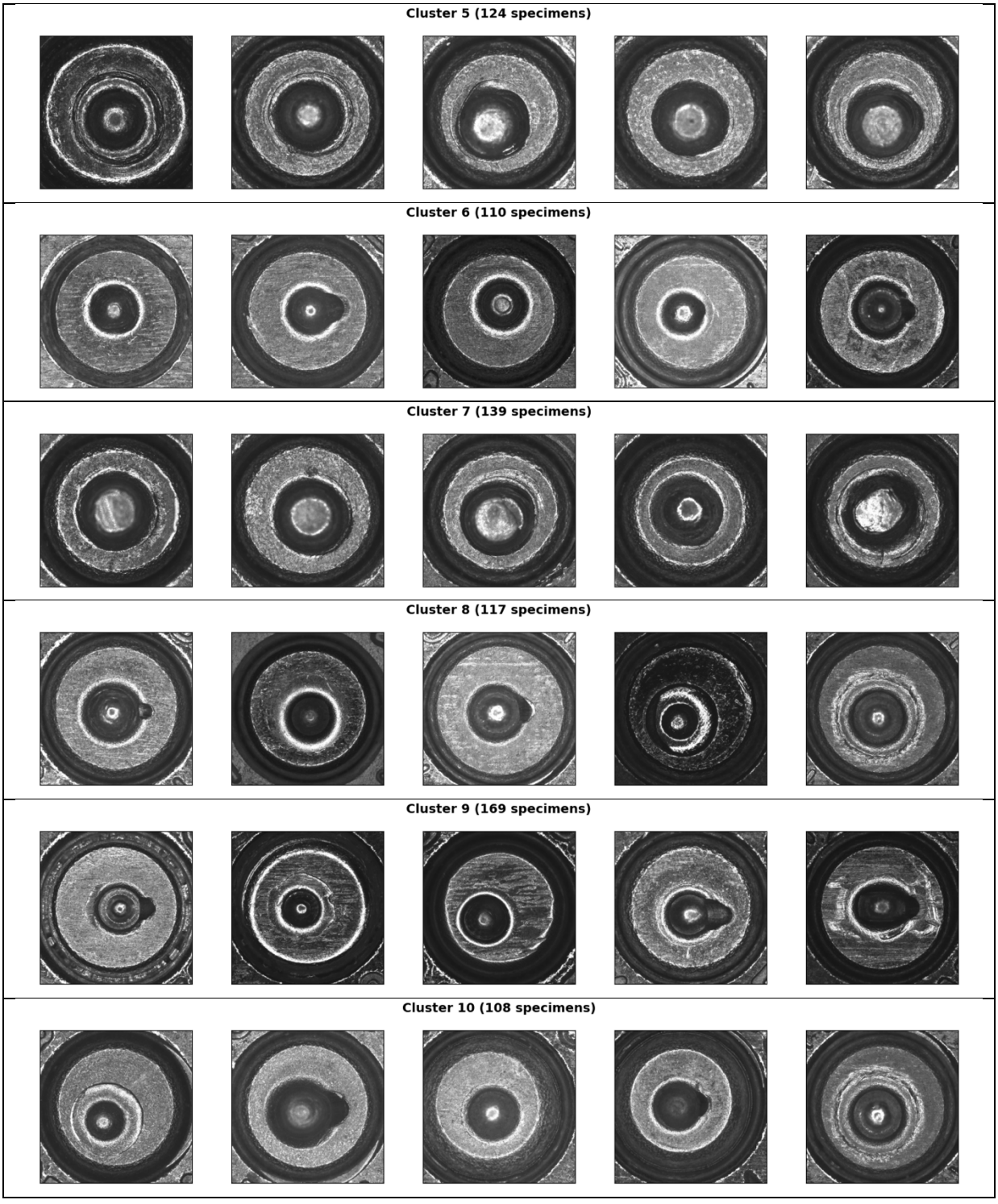
CONTRIBUTION #2 : K-MEANS

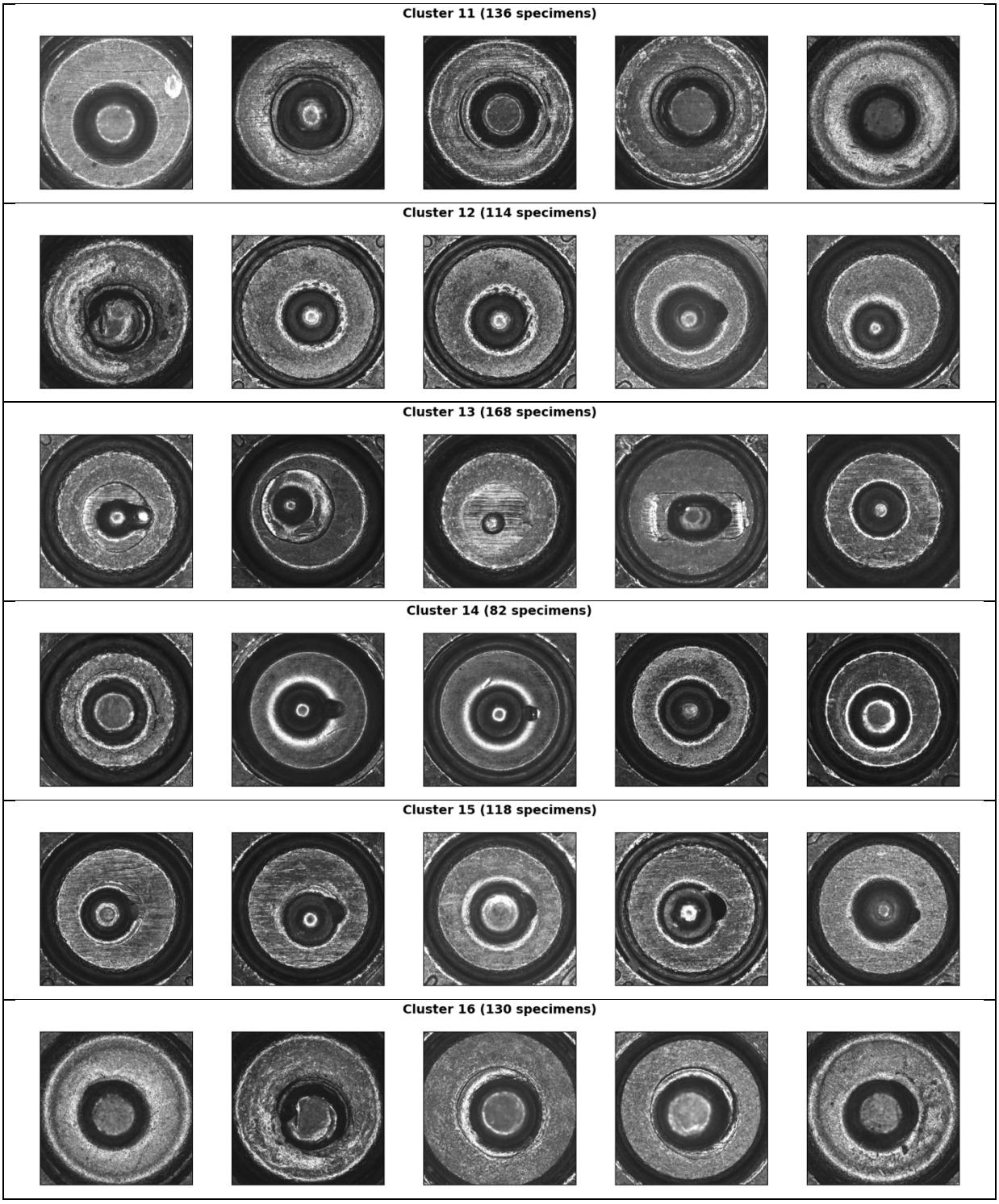
Images 2D des cinq premiers échantillons de chaque cluster K-means

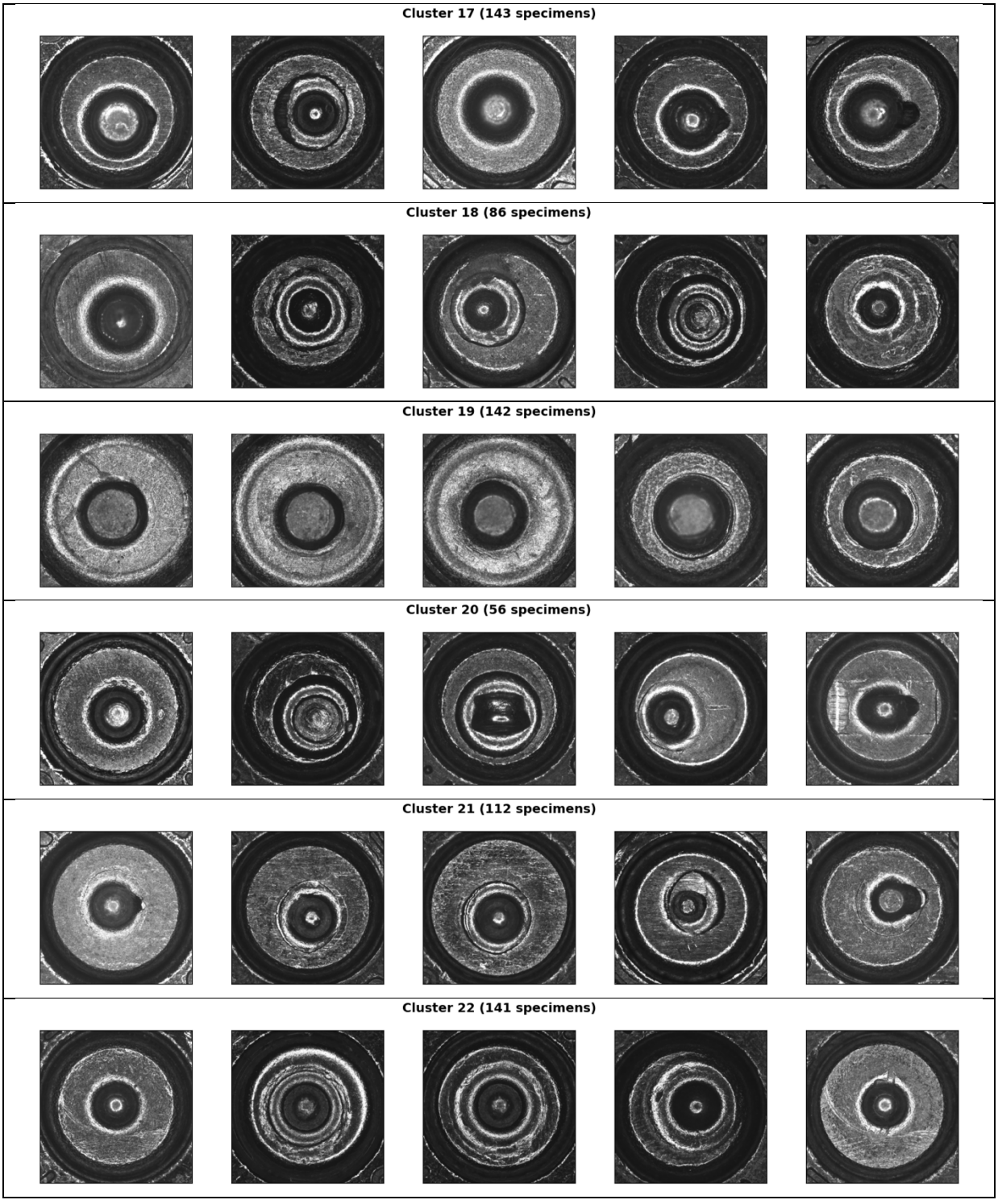
Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

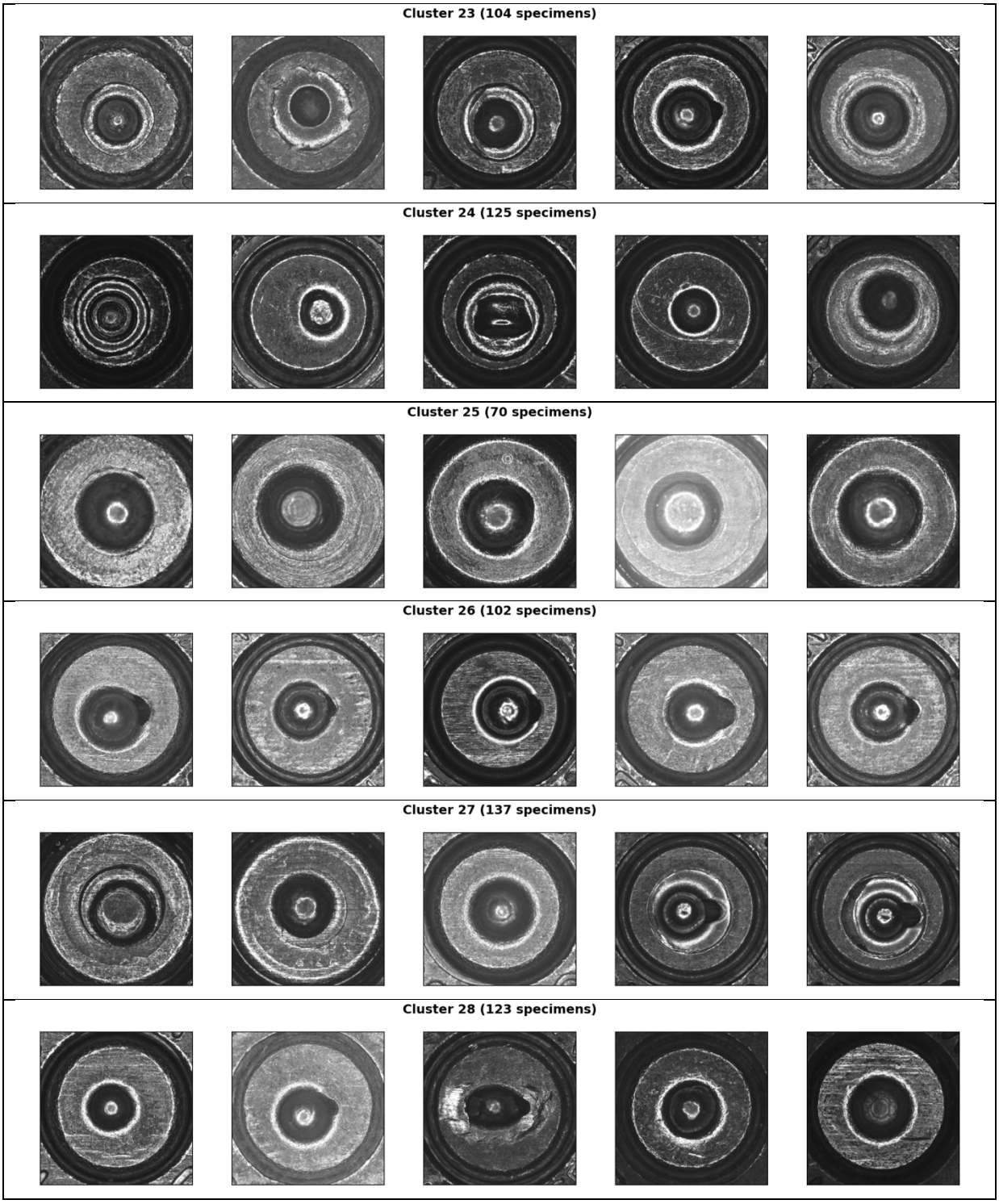
30 clusters




Cluster 5 (124 specimens)
Cluster 6 (110 specimens)
Cluster 7 (139 specimens)
Cluster 8 (117 specimens)
Cluster 9 (169 specimens)
Cluster 10 (108 specimens)


Cluster 11 (136 specimens)
Cluster 12 (114 specimens)
Cluster 13 (168 specimens)
Cluster 14 (82 specimens)
Cluster 15 (118 specimens)
Cluster 16 (130 specimens)


Cluster 17 (143 specimens)
Cluster 18 (86 specimens)
Cluster 19 (142 specimens)
Cluster 20 (56 specimens)
Cluster 21 (112 specimens)
Cluster 22 (141 specimens)


Cluster 23 (104 specimens)
Cluster 24 (125 specimens)
Cluster 25 (70 specimens)
Cluster 26 (102 specimens)
Cluster 27 (137 specimens)
Cluster 28 (123 specimens)

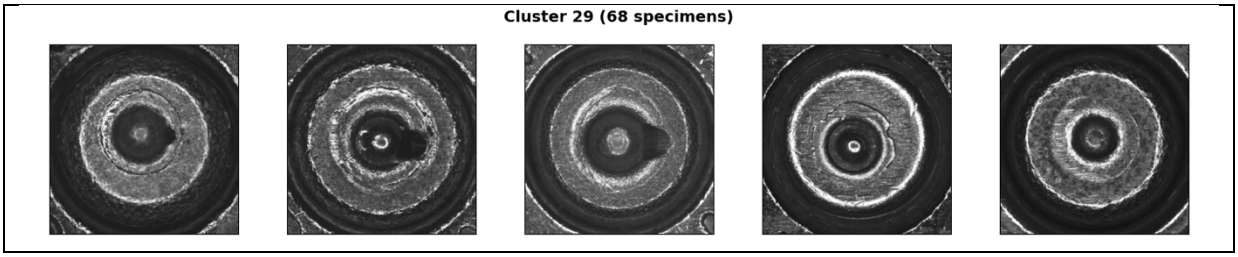
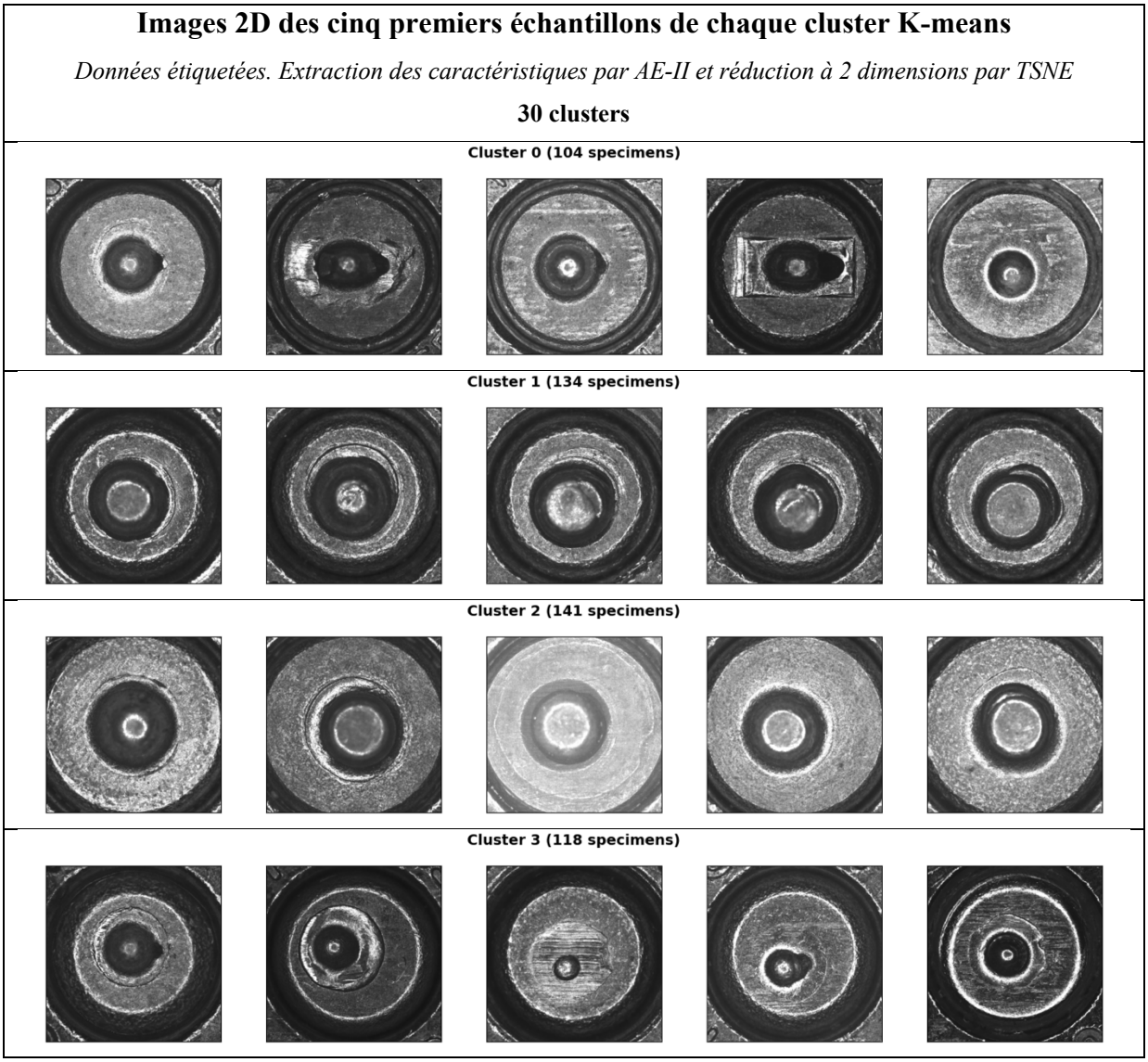
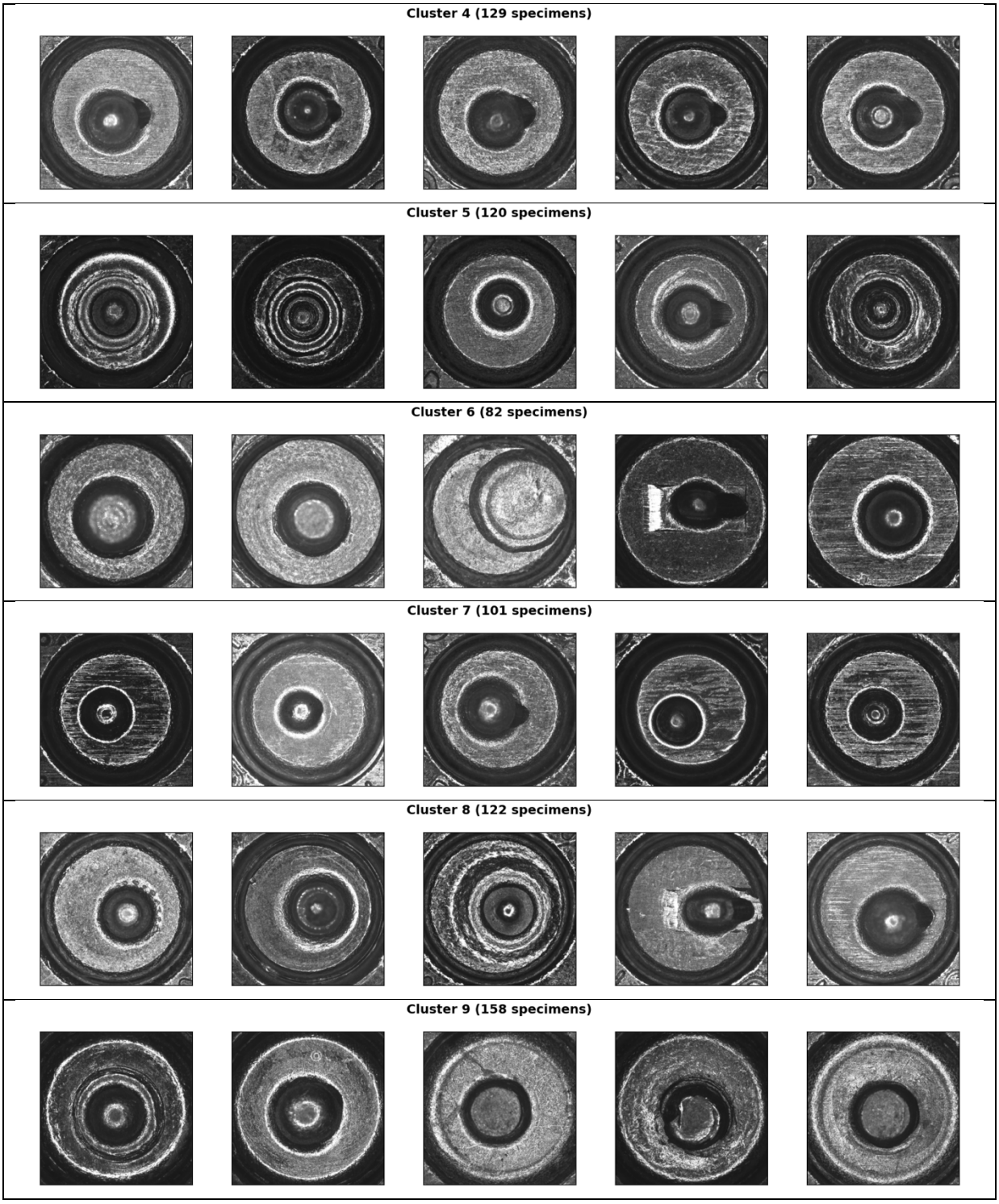
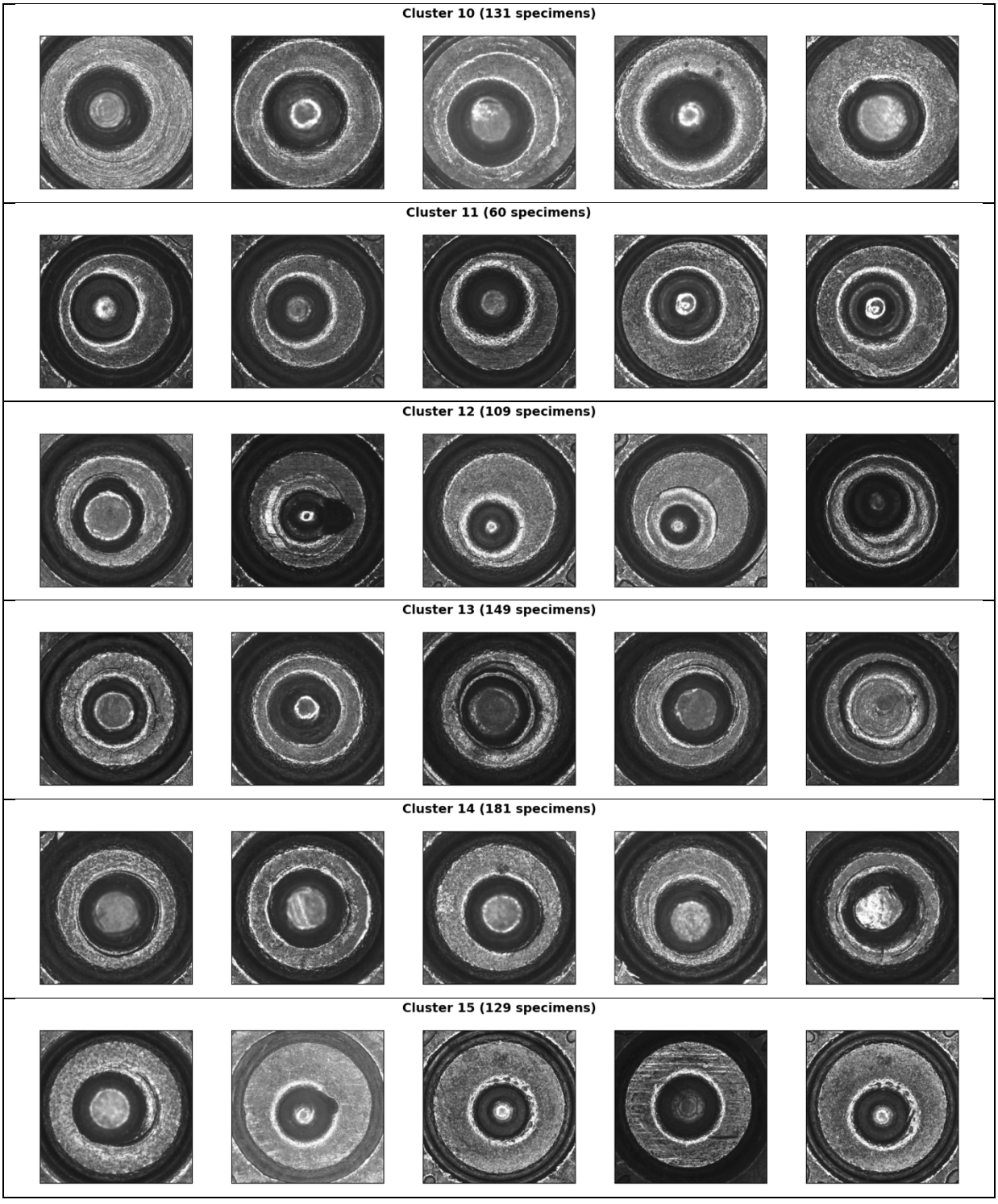
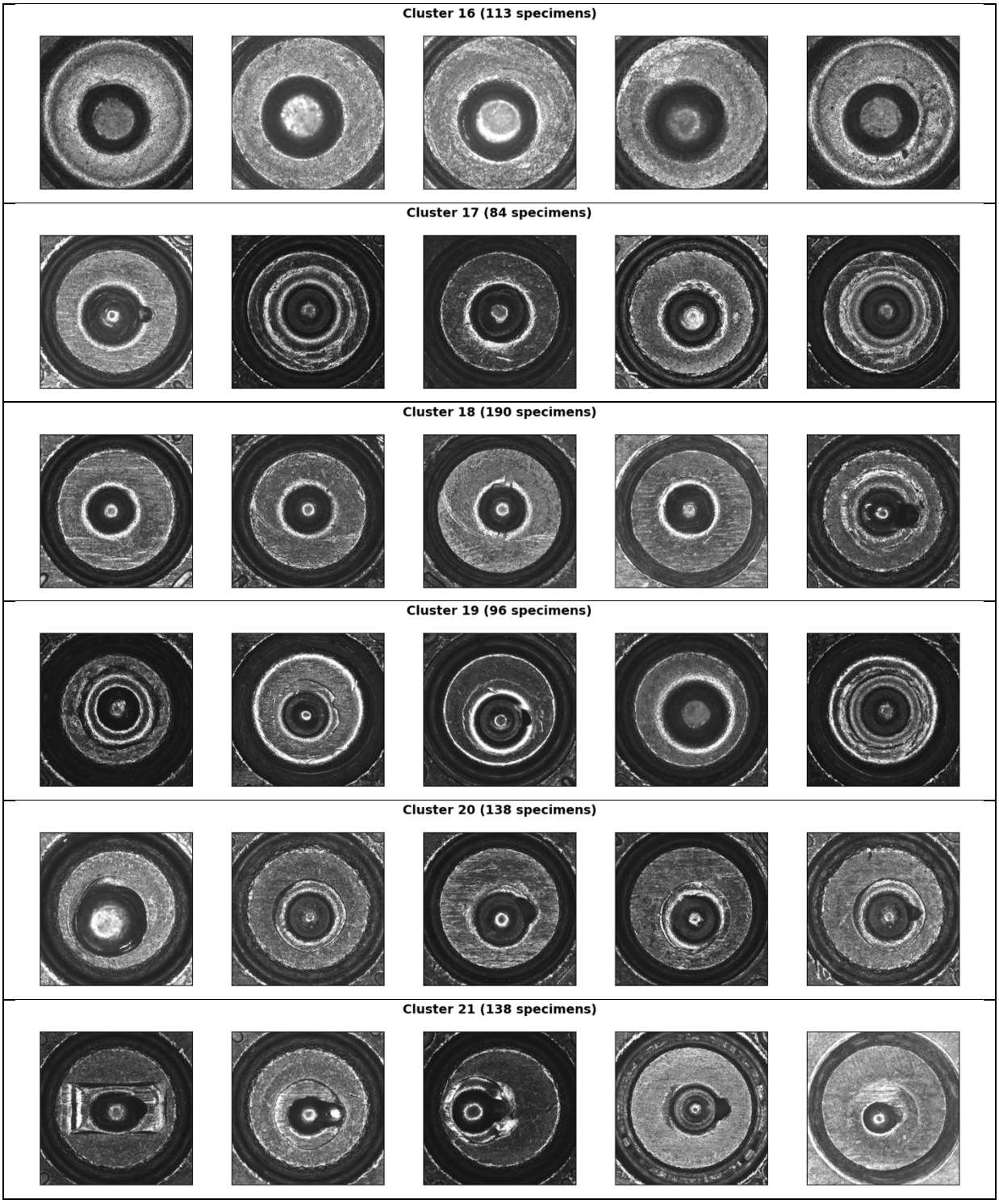


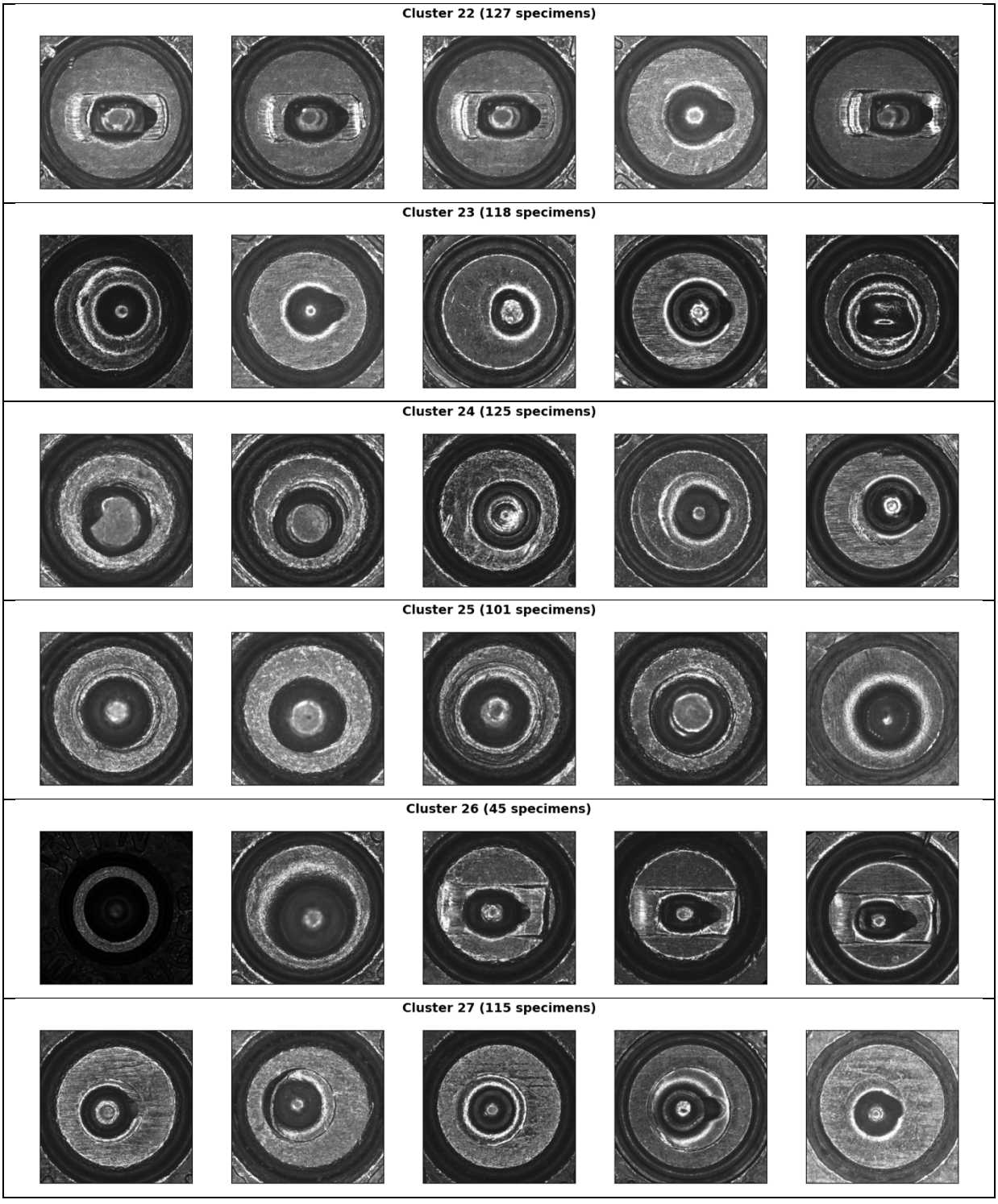
Figure-A IV-1 Images 2D des cinq premiers échantillons de chaque cluster K-means, pour un nombre de clusters de 30. Extraction des caractéristiques par la CNN ENB3 et réduction à 2 dimensions par TSNE




Cluster 4 (129 specimens)
Cluster 5 (120 specimens)
Cluster 6 (82 specimens)
Cluster 7 (101 specimens)
Cluster 8 (122 specimens)
Cluster 9 (158 specimens)


Cluster 10 (131 specimens)
Cluster 11 (60 specimens)
Cluster 12 (109 specimens)
Cluster 13 (149 specimens)
Cluster 14 (181 specimens)
Cluster 15 (129 specimens)


Cluster 16 (113 specimens)
Cluster 17 (84 specimens)
Cluster 18 (190 specimens)
Cluster 19 (96 specimens)
Cluster 20 (138 specimens)
Cluster 21 (138 specimens)


Cluster 22 (127 specimens)
Cluster 23 (118 specimens)
Cluster 24 (125 specimens)
Cluster 25 (101 specimens)
Cluster 26 (45 specimens)
Cluster 27 (115 specimens)

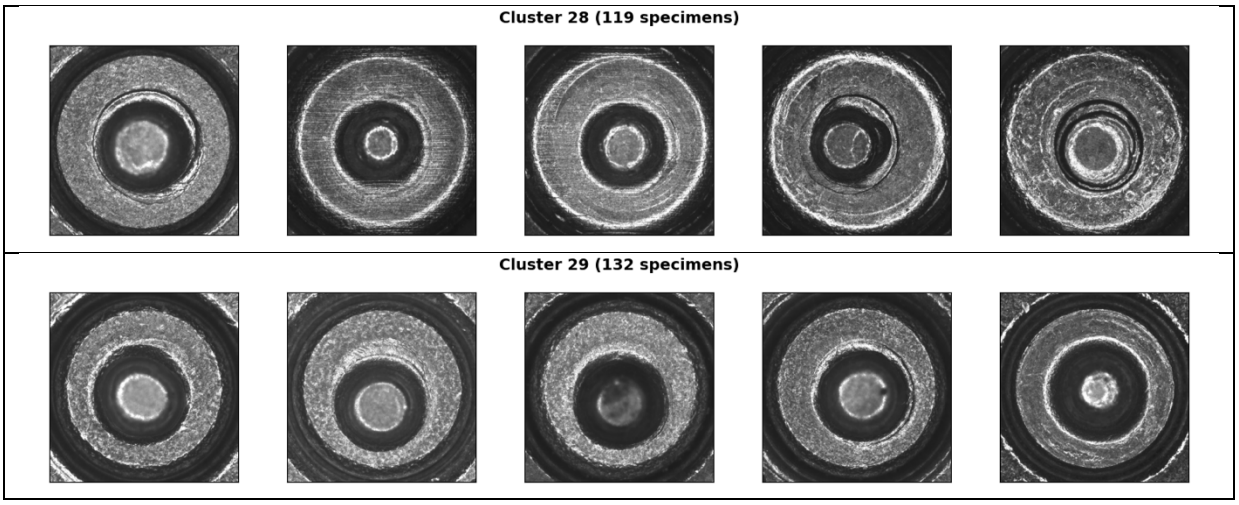


Figure-A IV-2 Images 2D des cinq premiers échantillons de chaque cluster K-means, pour un nombre de clusters de 30. Extraction des caractéristiques par l'encodeur de l'autoencodeur AE-II et réduction à 2 dimensions par TSN

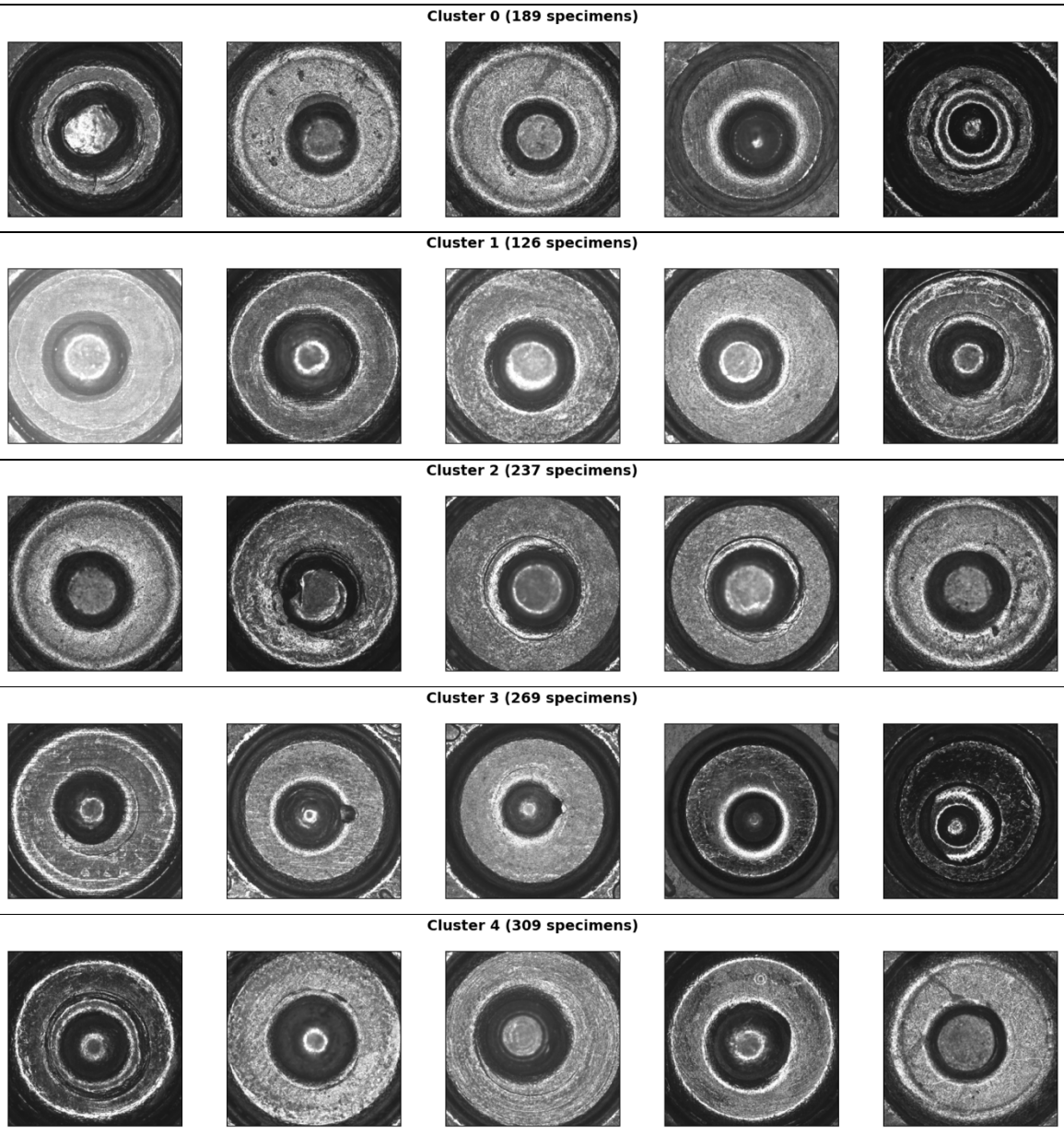
ANNEXE V

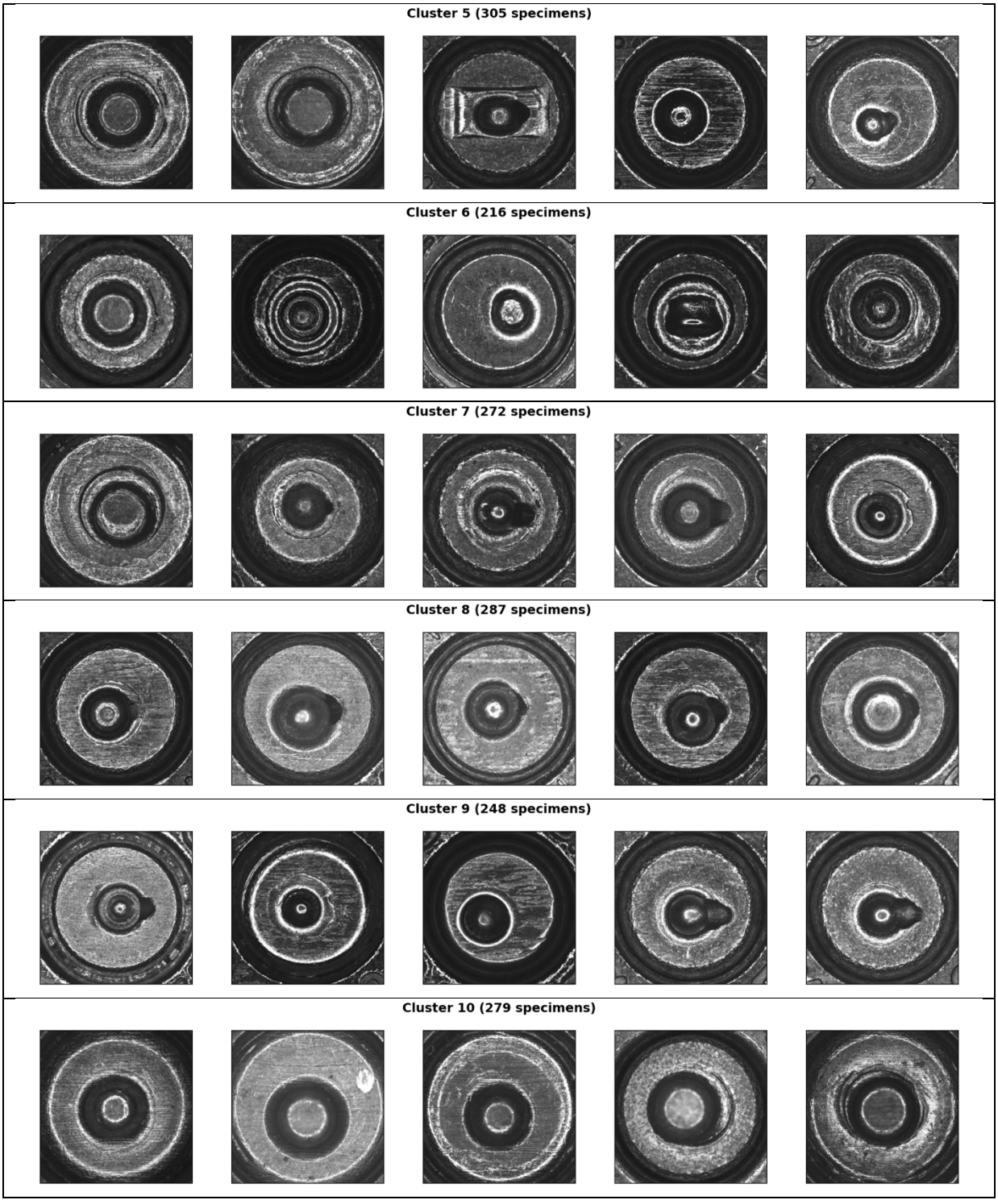
CONTRIBUTION #2 : FUZZY C-MEANS

Images 2D des cinq premiers échantillons de chaque cluster Fuzzy C-means

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

15 clusters




Cluster 5 (305 specimens)
Cluster 6 (216 specimens)
Cluster 7 (272 specimens)
Cluster 8 (287 specimens)
Cluster 9 (248 specimens)
Cluster 10 (279 specimens)

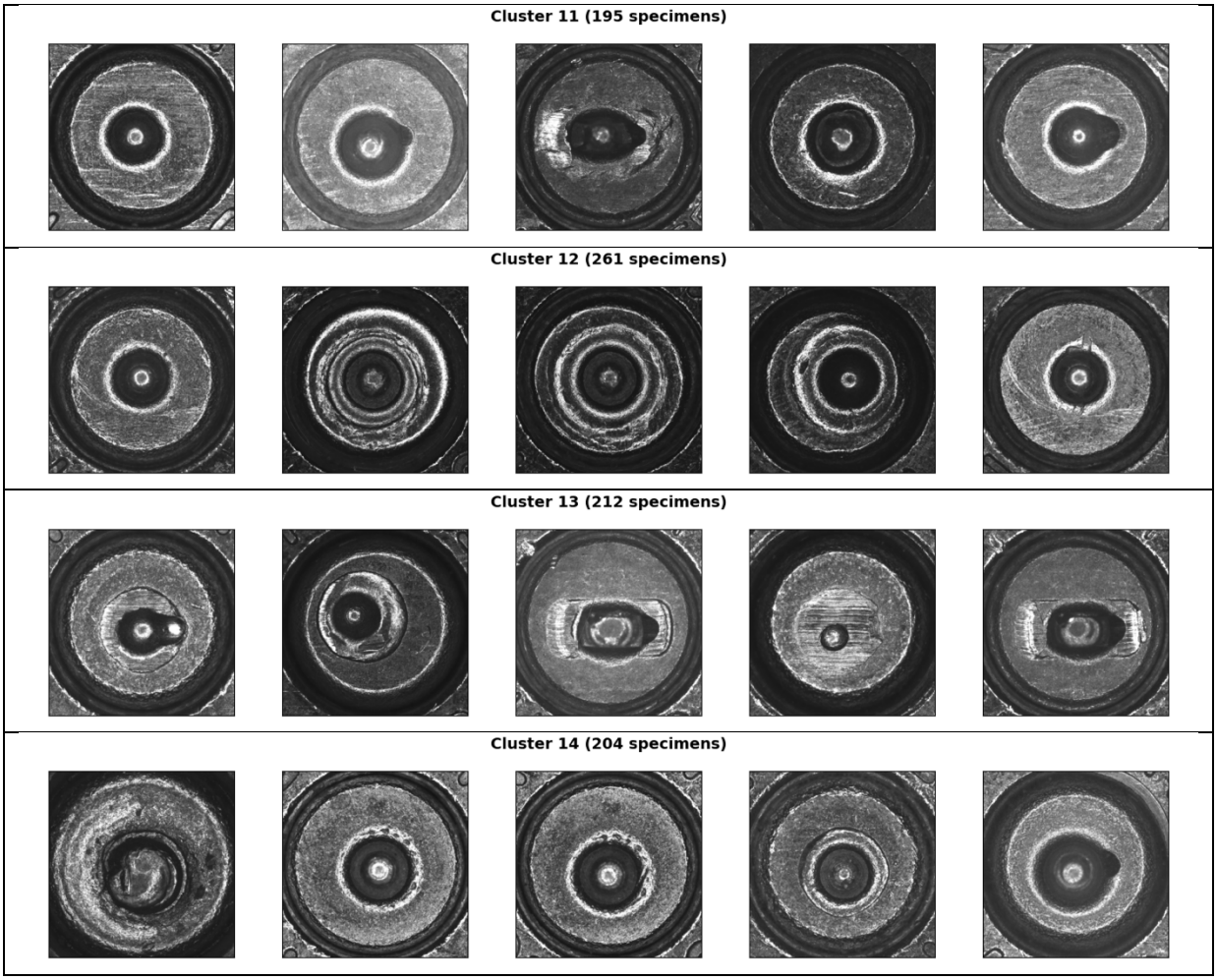


Figure-A V-1 Images 2D des cinq premiers échantillons de chaque cluster Fuzzy C-means, pour un nombre de clusters de 15. Extraction des caractéristiques par la CNN ENB3 et réduction à 2 dimensions par TSNE

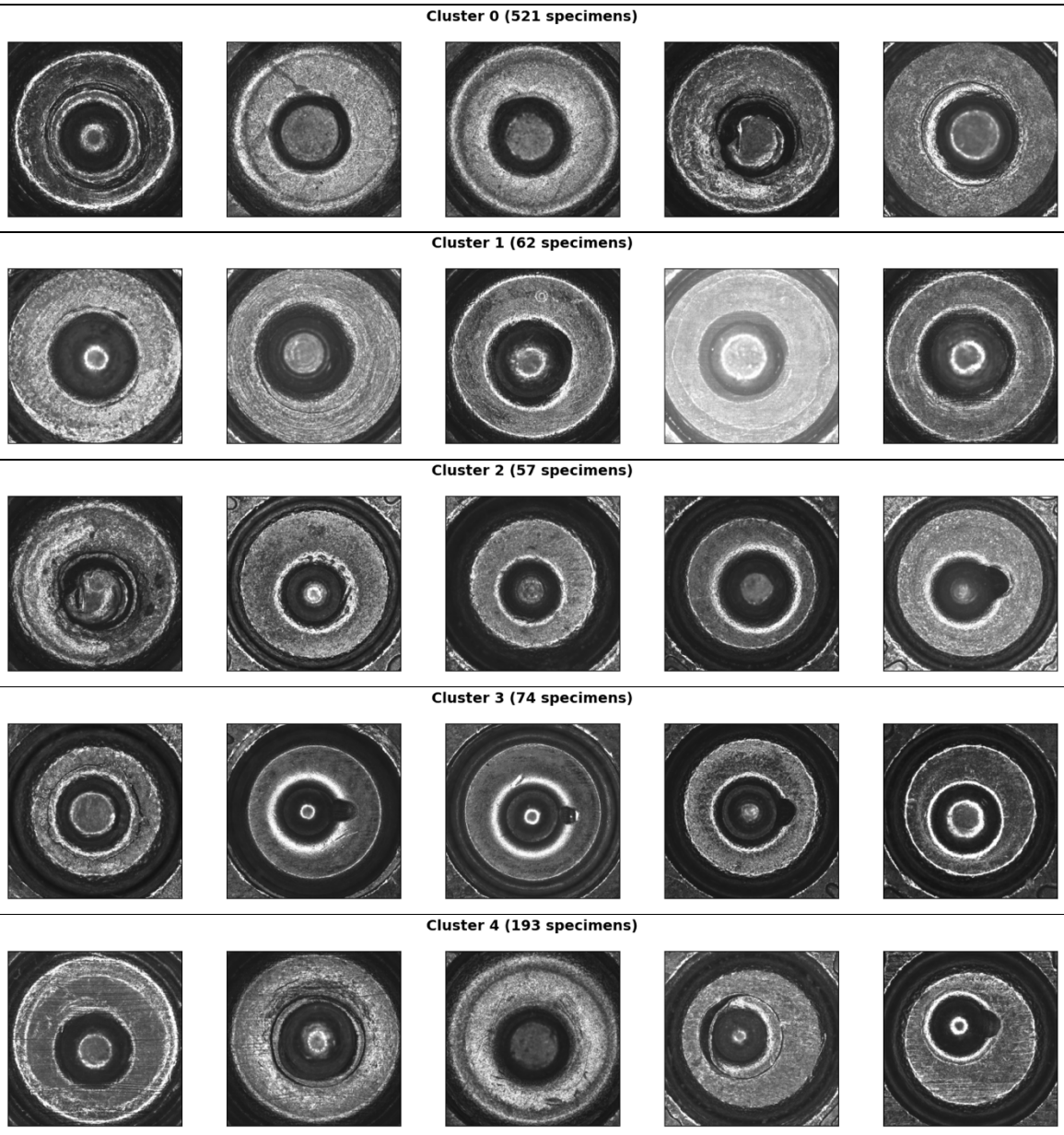
ANNEXE VI

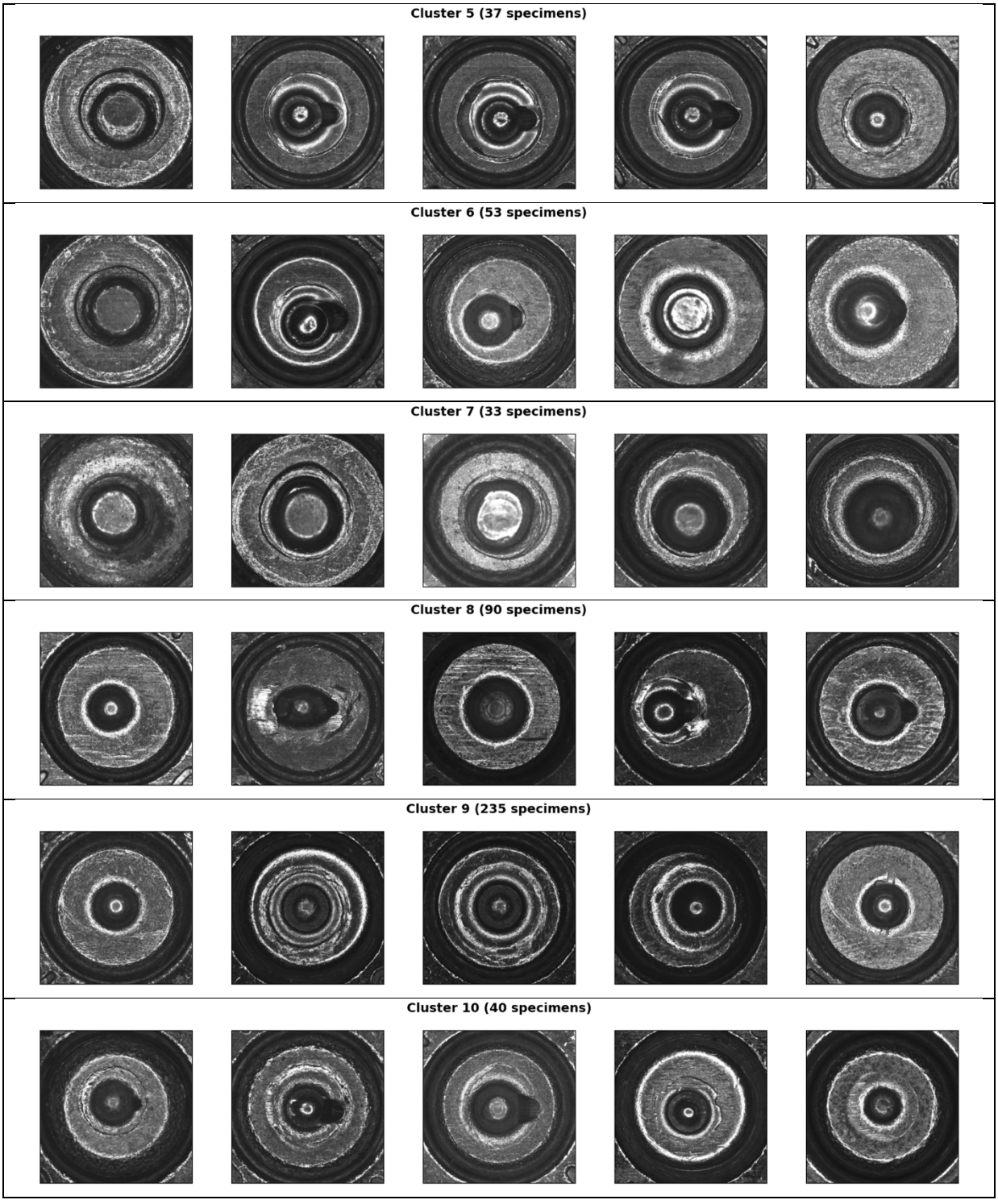
CONTRIBUTION #2 : DBSCAN

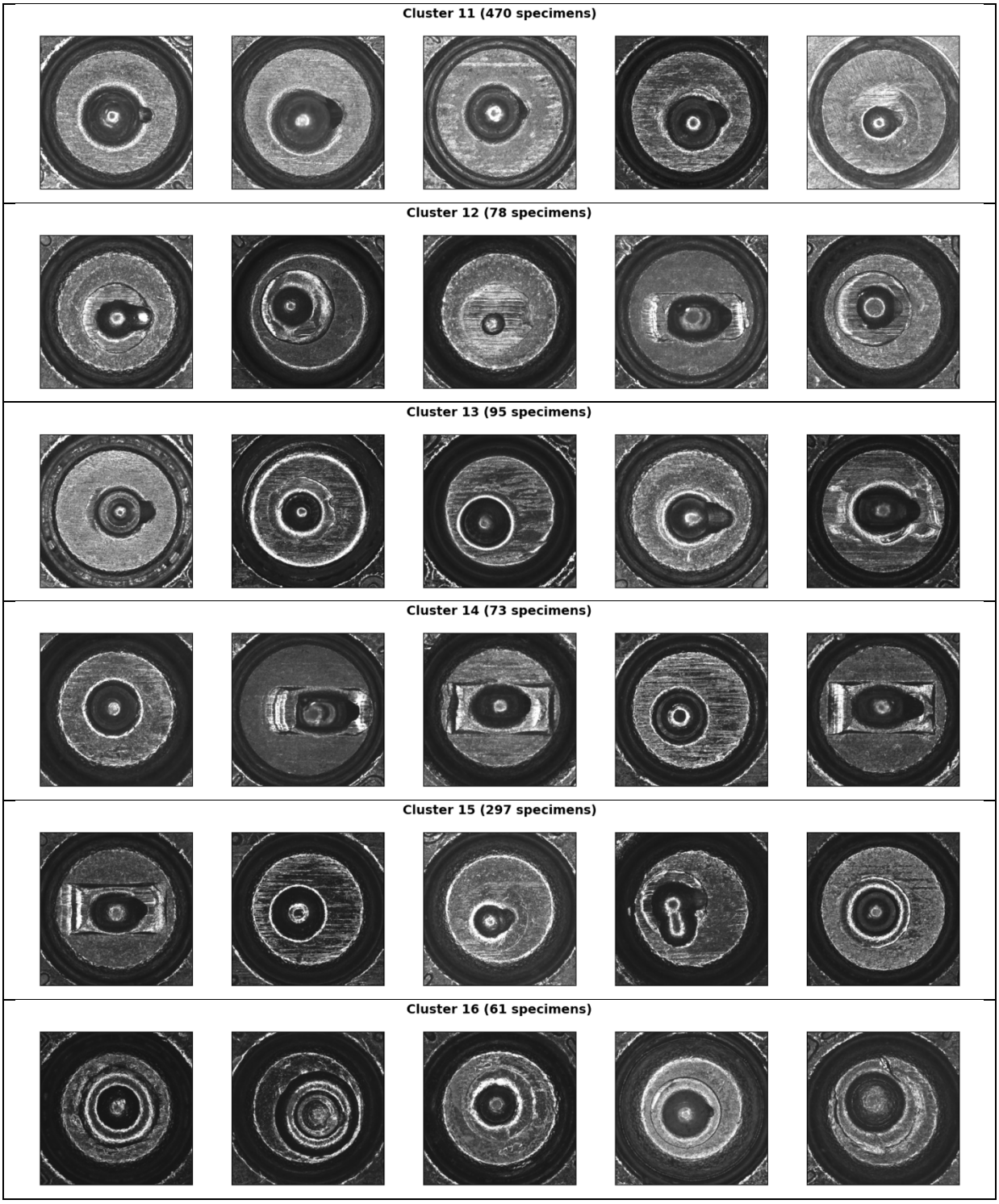
Images 2D des cinq premiers échantillons de chaque cluster DBSCAN

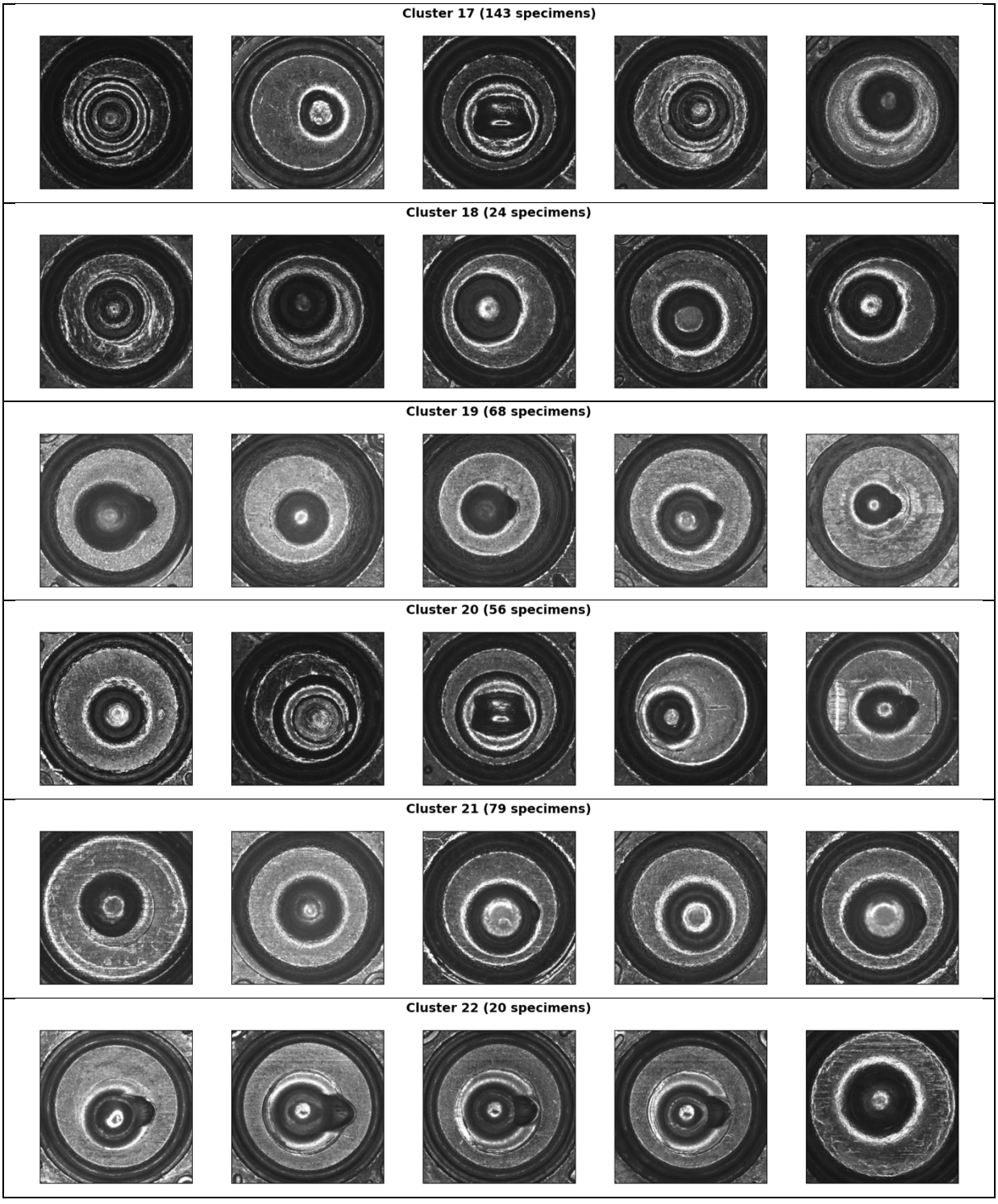
Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

29 clusters




Cluster 5 (37 specimens)
Cluster 6 (53 specimens)
Cluster 7 (33 specimens)
Cluster 8 (90 specimens)
Cluster 9 (235 specimens)
Cluster 10 (40 specimens)


Cluster 11 (470 specimens)
Cluster 12 (78 specimens)
Cluster 13 (95 specimens)
Cluster 14 (73 specimens)
Cluster 15 (297 specimens)
Cluster 16 (61 specimens)


Cluster 17 (143 specimens)
Cluster 18 (24 specimens)
Cluster 19 (68 specimens)
Cluster 20 (56 specimens)
Cluster 21 (79 specimens)
Cluster 22 (20 specimens)

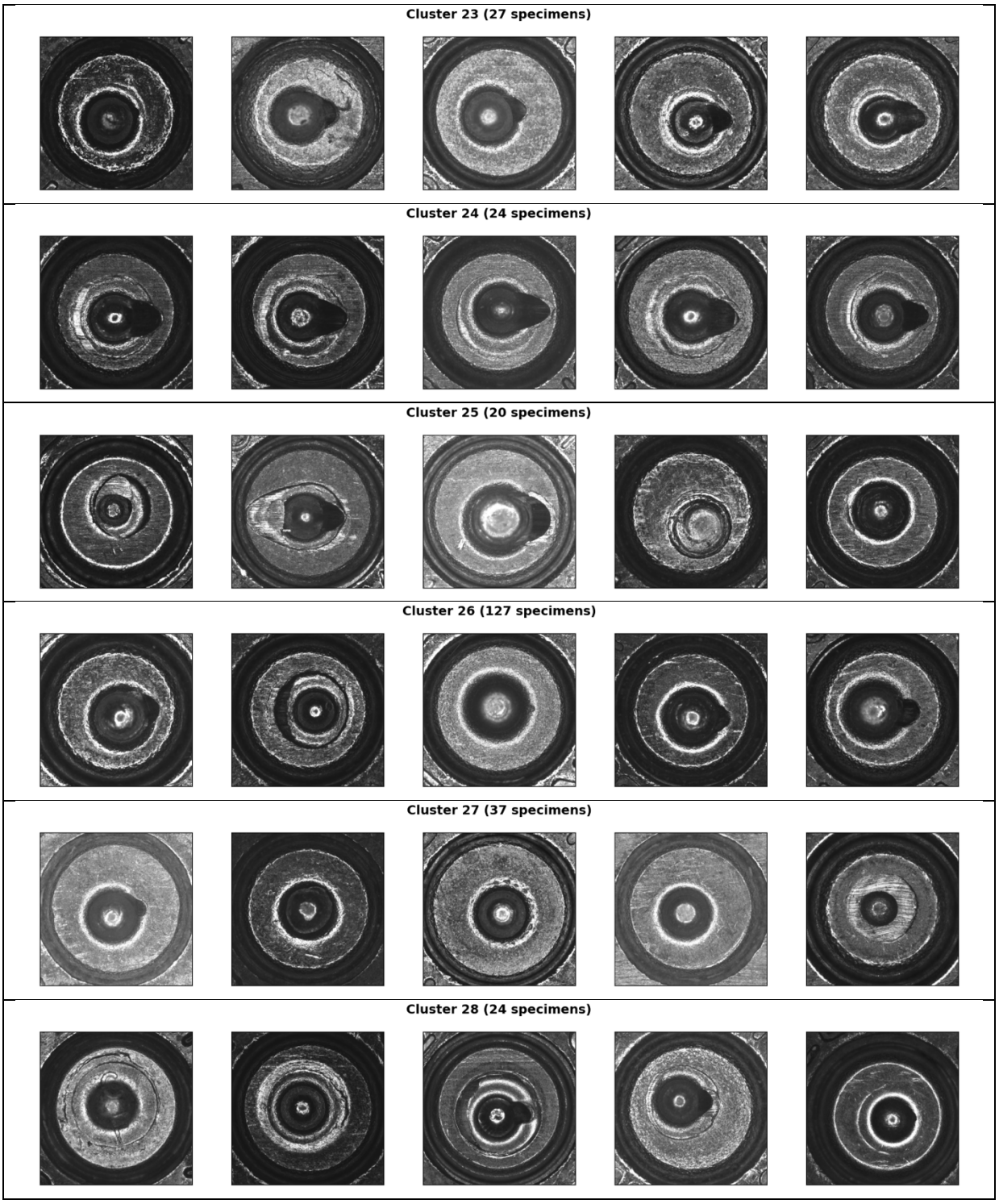


Figure-A VI-1 Images 2D des cinq premiers échantillons de chaque cluster DBSCAN, pour un nombre de clusters de 29. Extraction des caractéristiques par la CNN ENB3 et réduction à 2 dimensions par TSNE

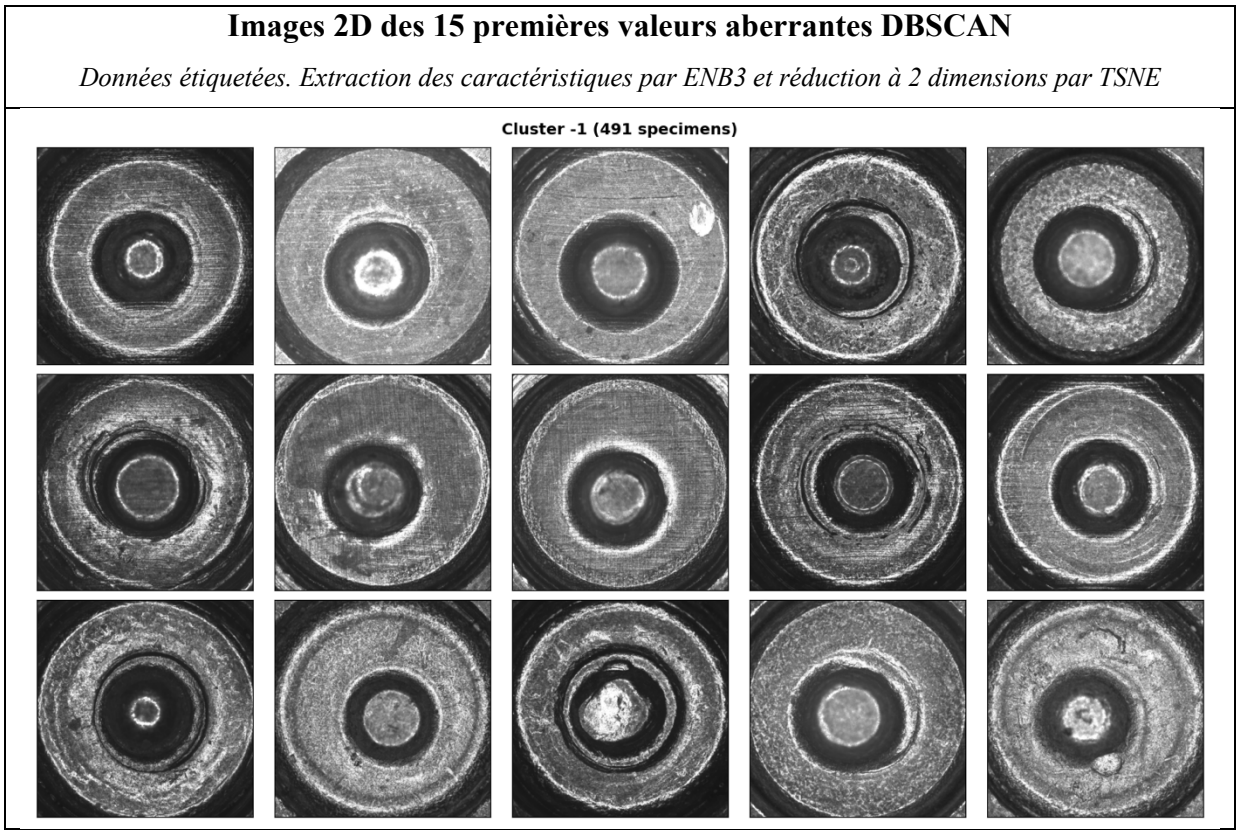


Figure-A VI-2 Images 2D des 15 premiers échantillons considérés comme étant des valeurs aberrantes par DBSCAN

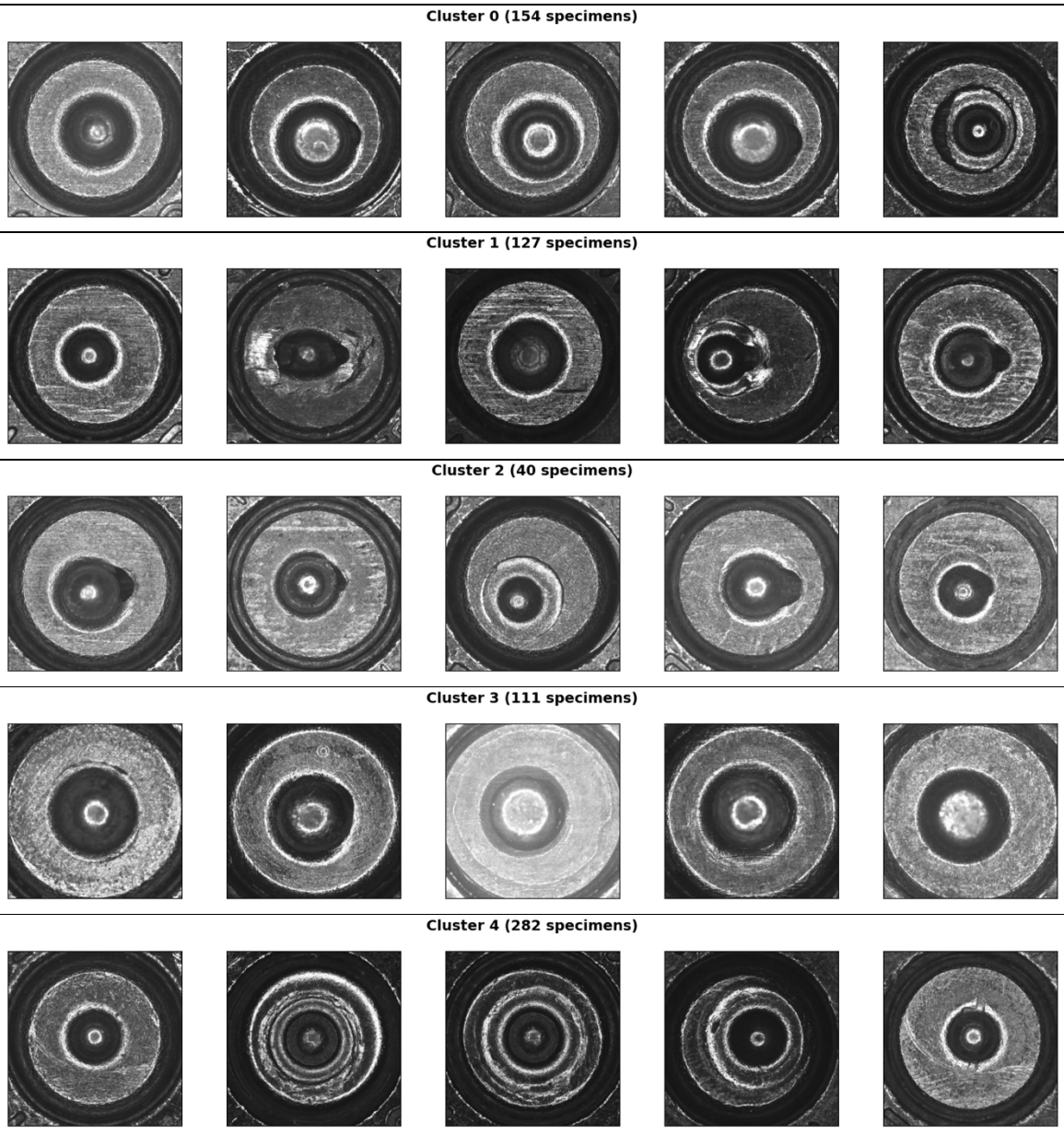
ANNEXE VII

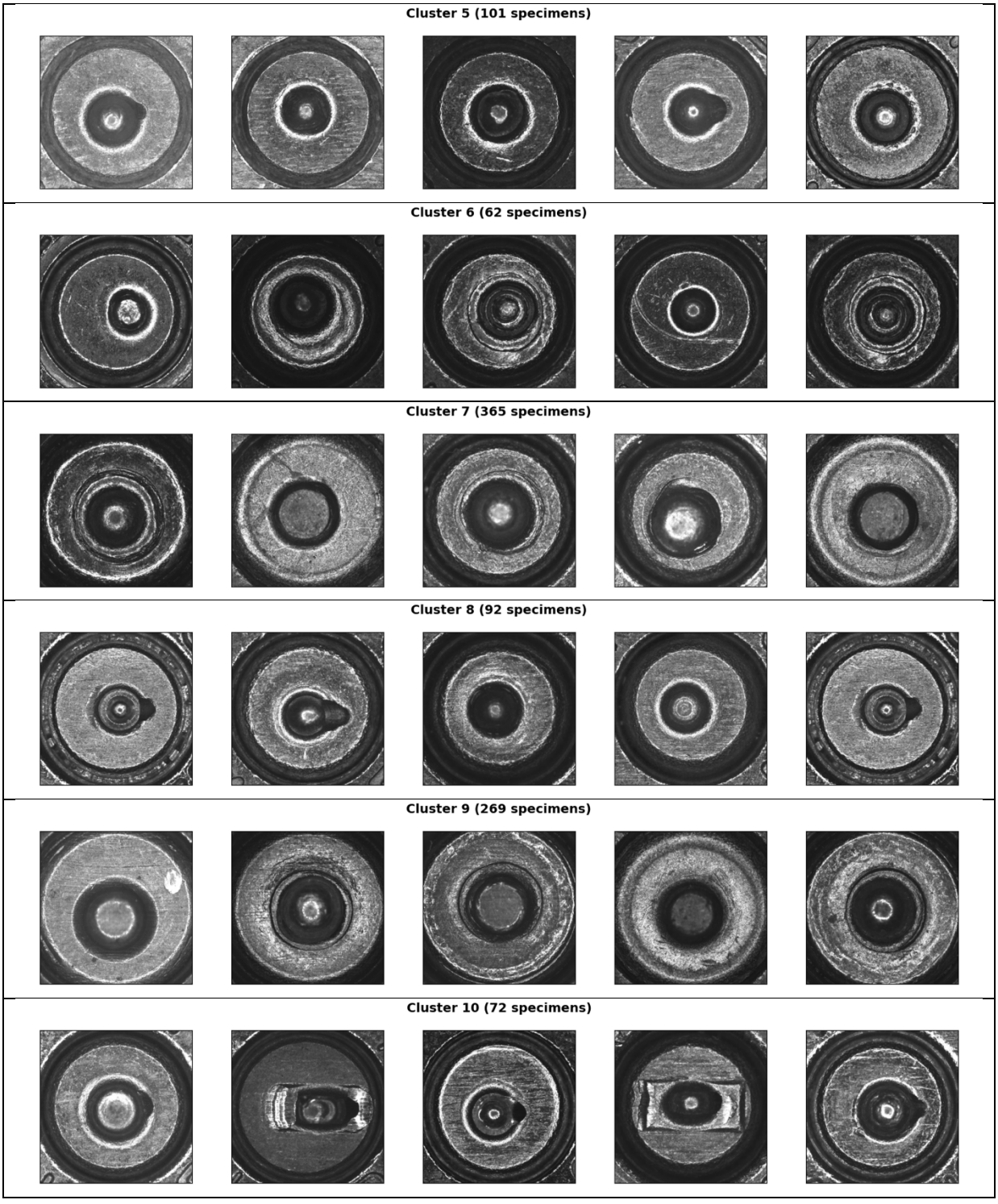
CONTRIBUTION #2 : HDBSCAN

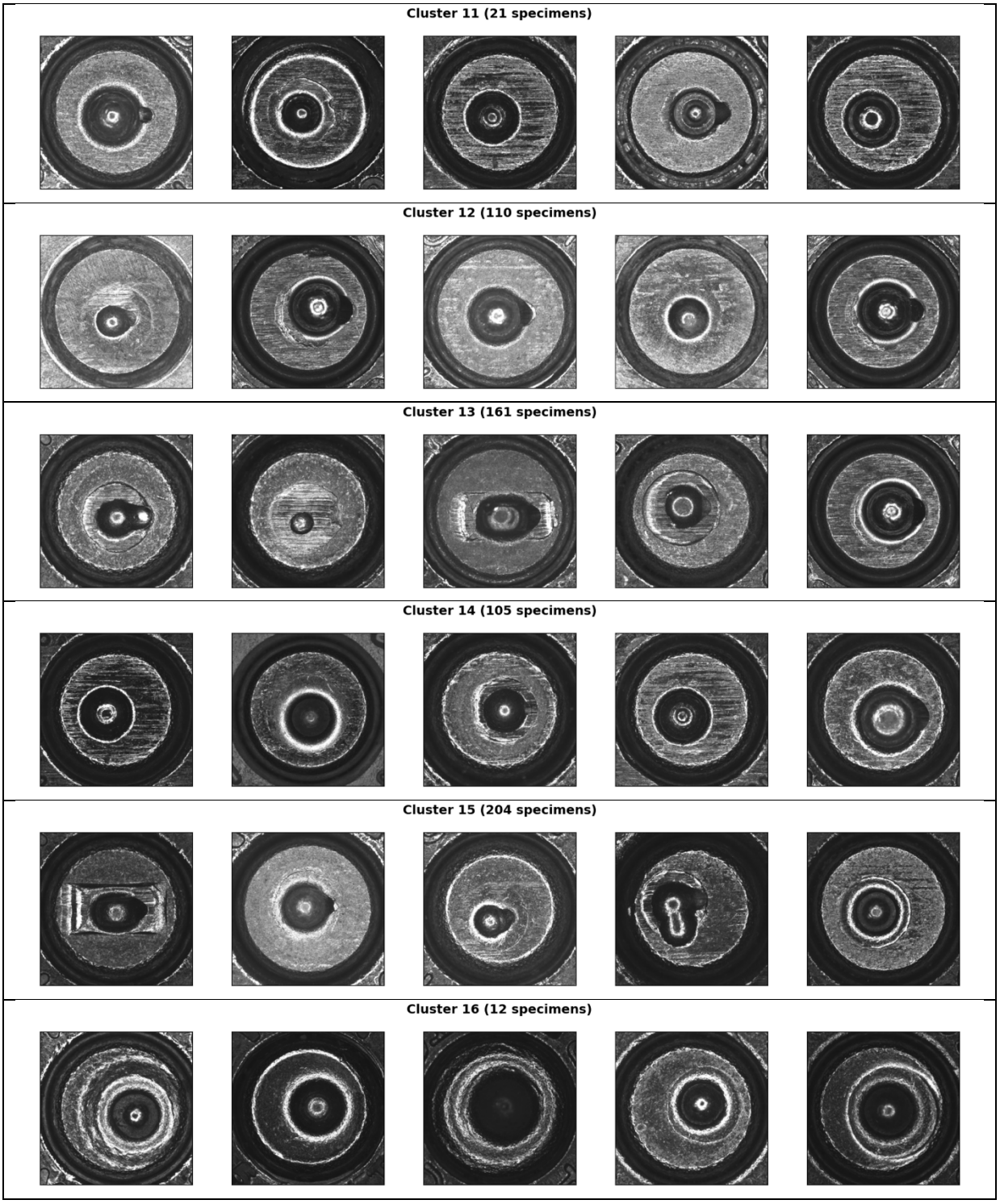
Images 2D des cinq premiers échantillons de chaque cluster HDBSCAN

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

19 clusters




Cluster 5 (101 specimens)
Cluster 6 (62 specimens)
Cluster 7 (365 specimens)
Cluster 8 (92 specimens)
Cluster 9 (269 specimens)
Cluster 10 (72 specimens)


Cluster 11 (21 specimens)
Cluster 12 (110 specimens)
Cluster 13 (161 specimens)
Cluster 14 (105 specimens)
Cluster 15 (204 specimens)
Cluster 16 (12 specimens)

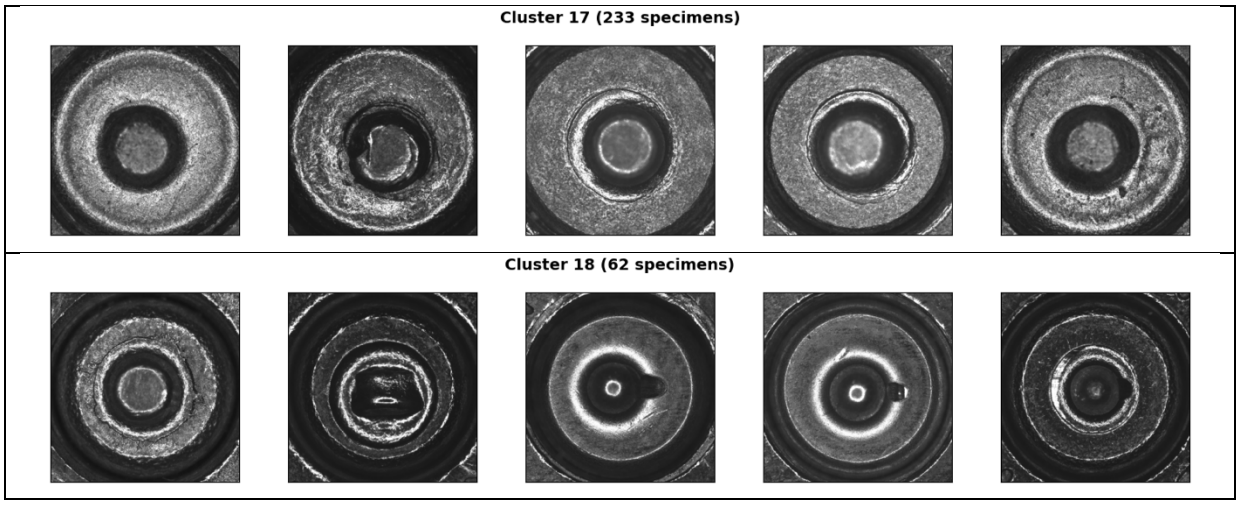


Figure-A VII-1 Images 2D des cinq premiers échantillons de chaque cluster HDBSCAN, pour un nombre de clusters de 19. Extraction des caractéristiques par la CNN ENB3 et réduction à 2 dimensions par TSNE

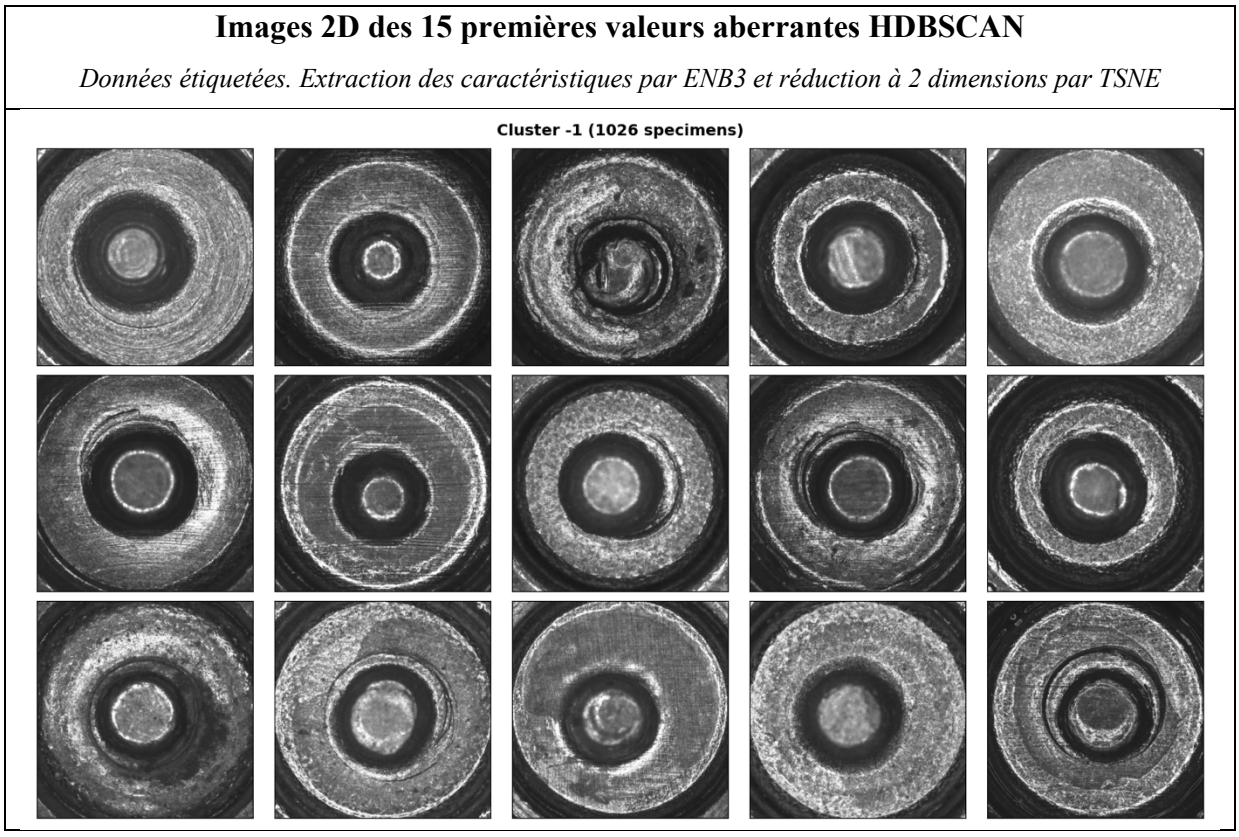


Figure-A VII-2 Images 2D des 15 premiers échantillons considérés comme étant des valeurs aberrantes par HDBSCAN

ANNEXE VIII

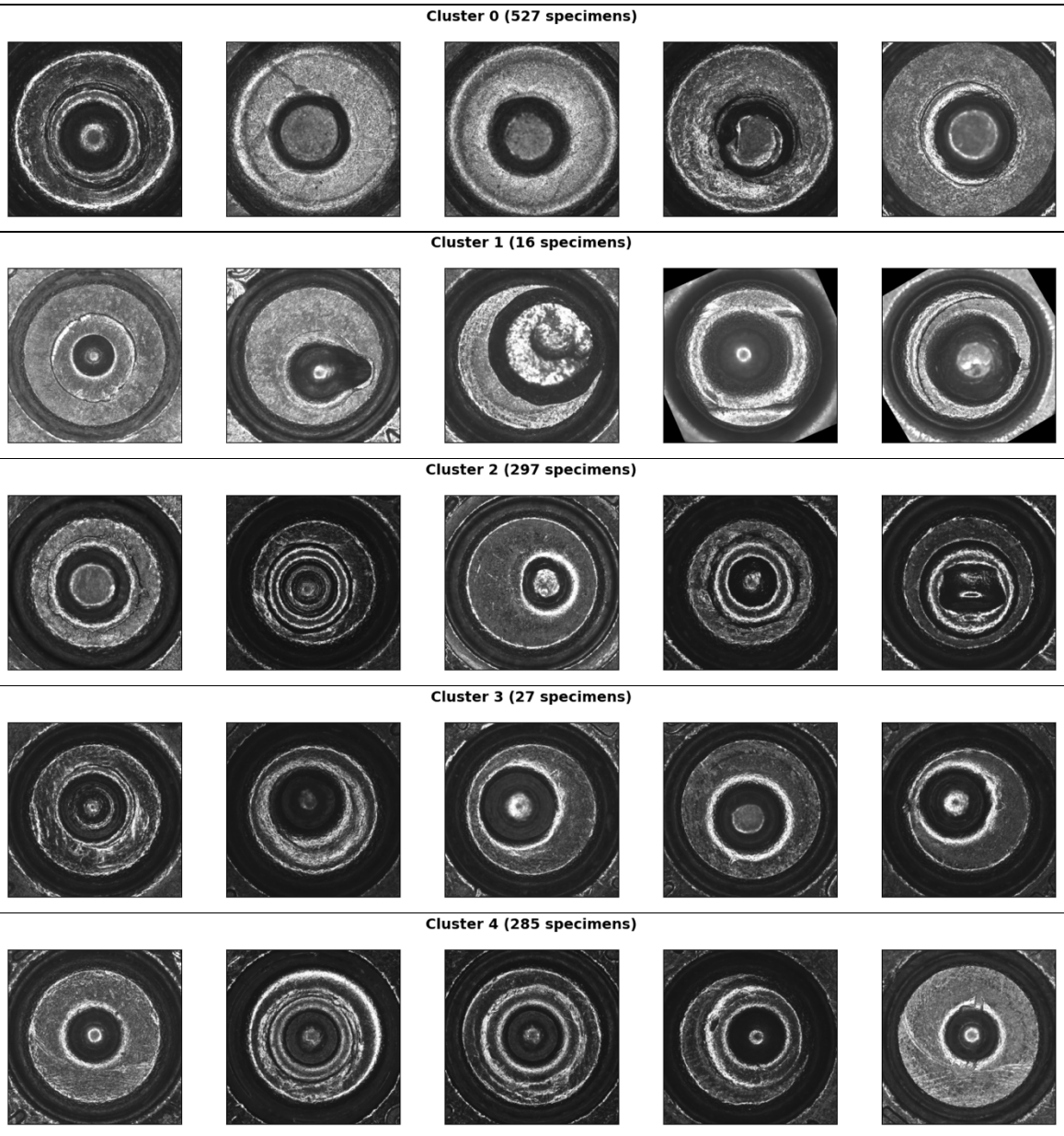
CONTRIBUTION #2 : OPTICS

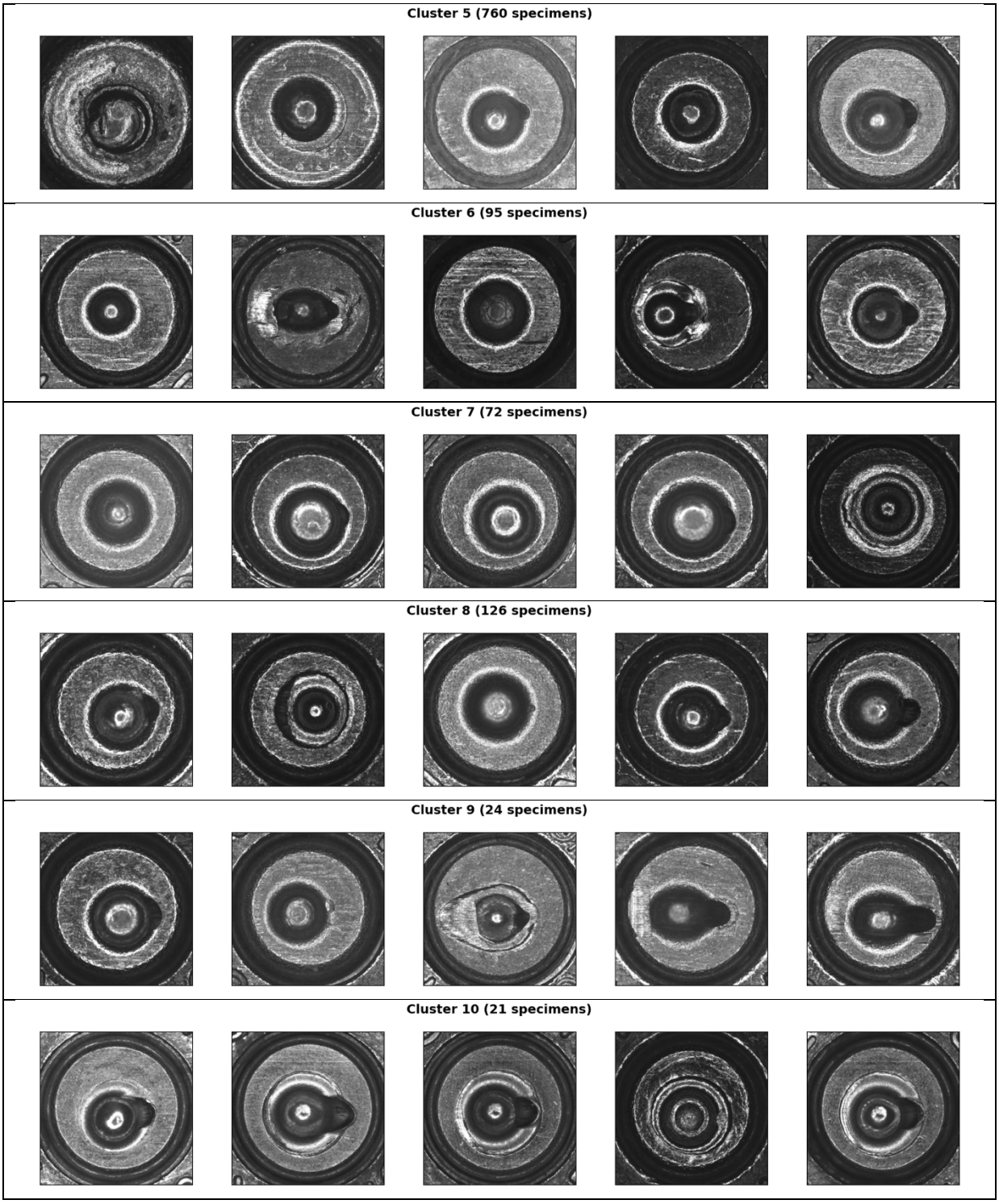
Images 2D des cinq premiers échantillons de chaque cluster OPTICS

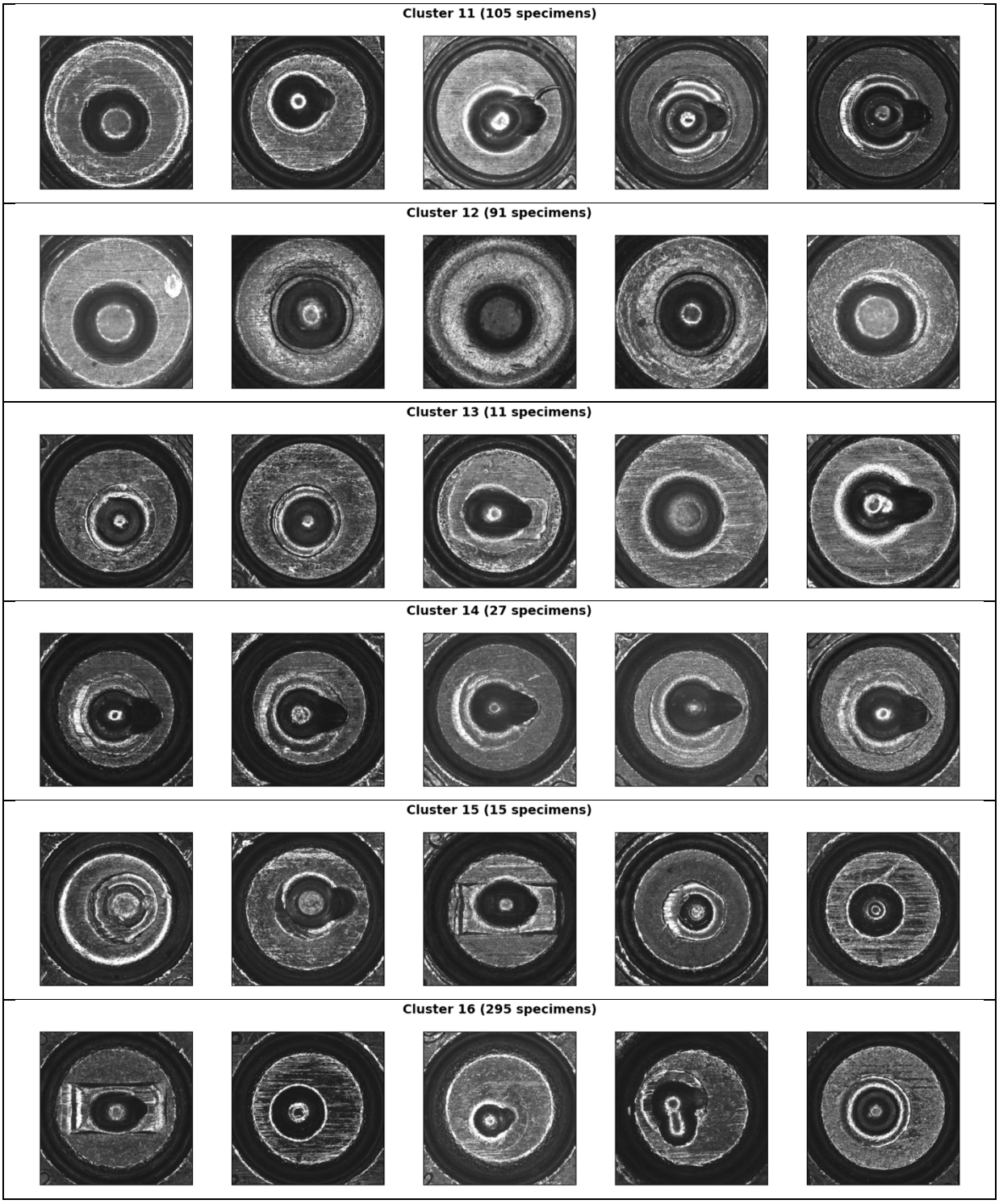
Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

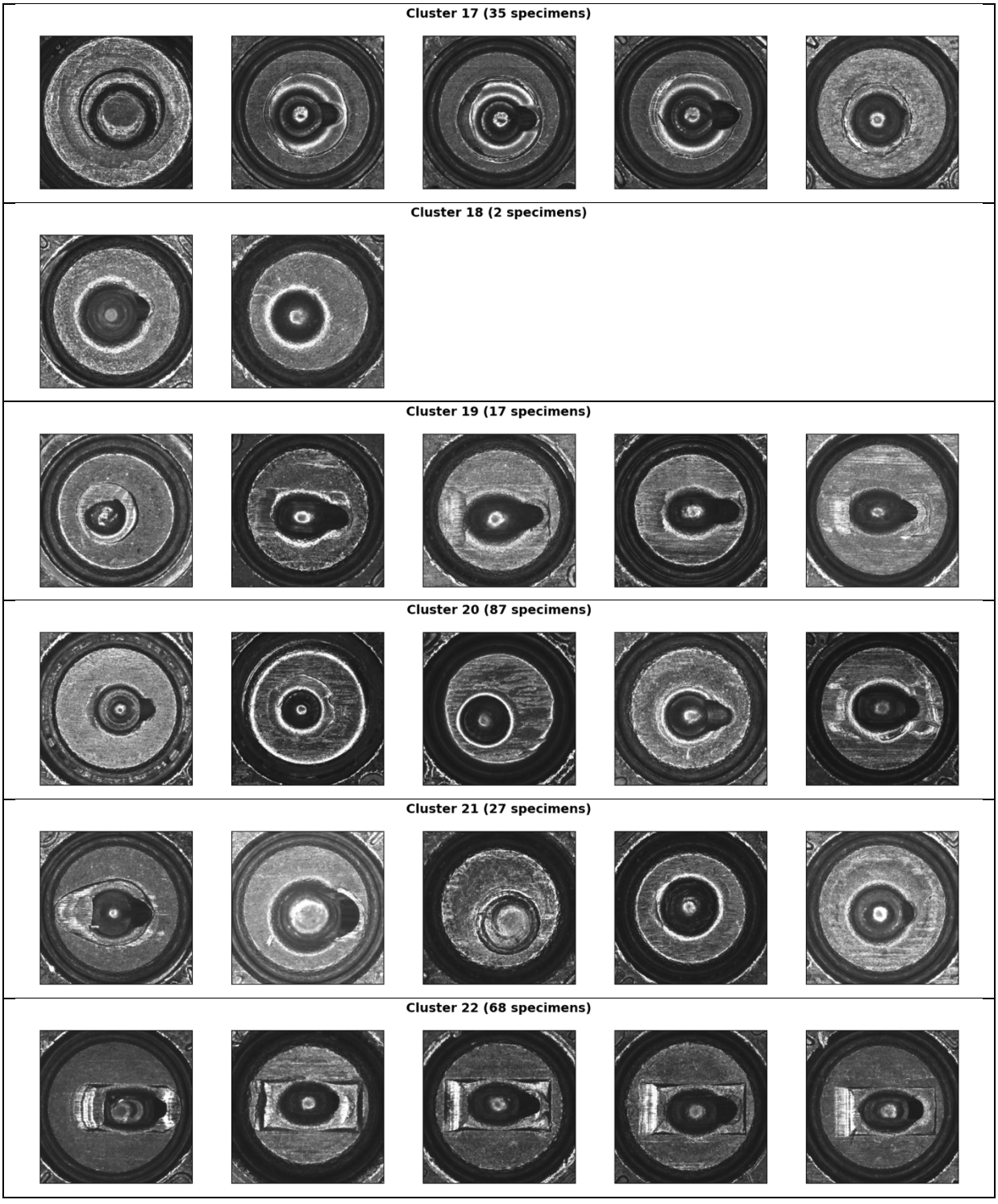
Paramètres : ***métrique de distance : Minkowski, min_cluster_size : 5, min_samples : 15***

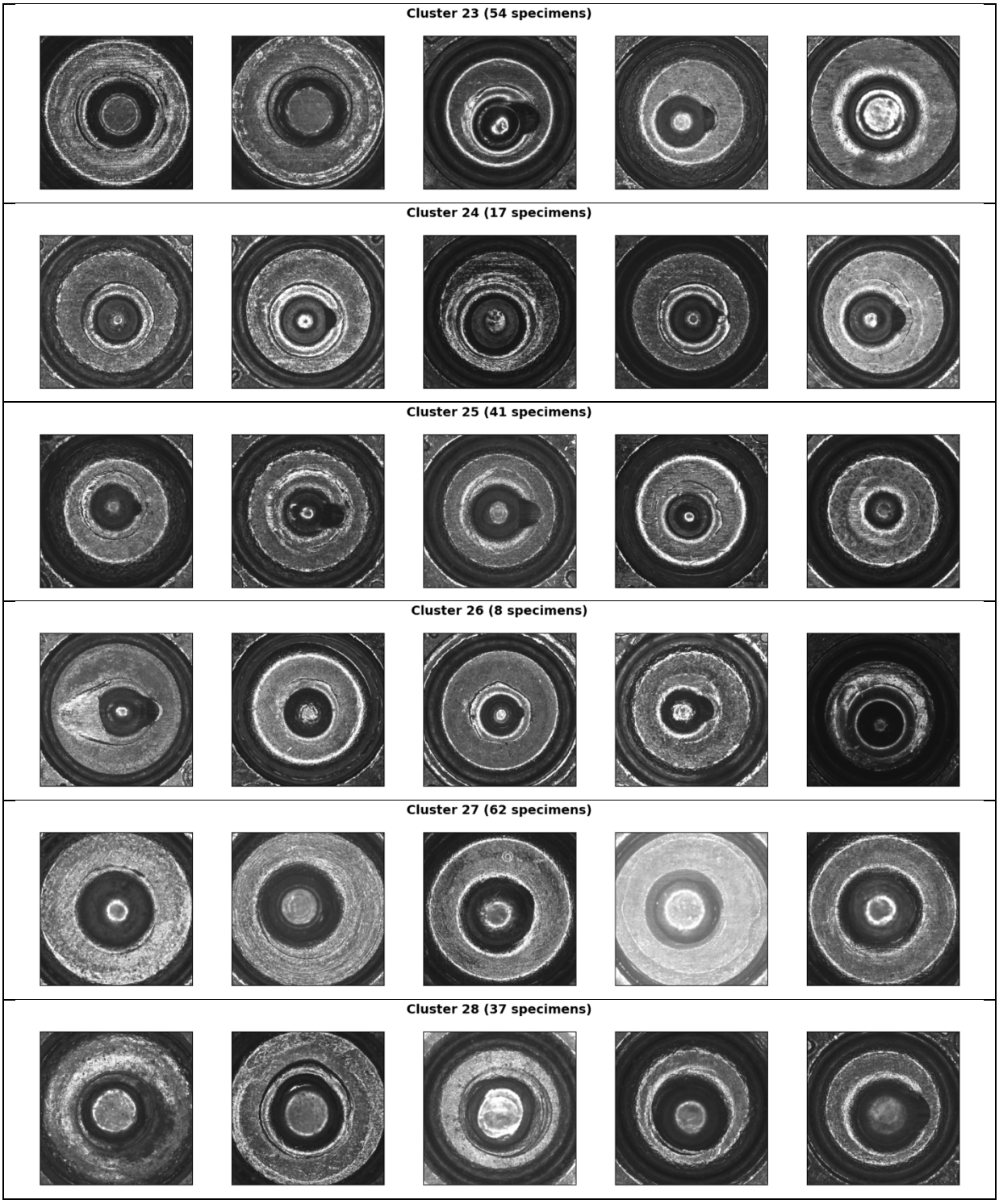
30 clusters




Cluster 5 (760 specimens)
Cluster 6 (95 specimens)
Cluster 7 (72 specimens)
Cluster 8 (126 specimens)
Cluster 9 (24 specimens)
Cluster 10 (21 specimens)


Cluster 11 (105 specimens)
Cluster 12 (91 specimens)
Cluster 13 (11 specimens)
Cluster 14 (27 specimens)
Cluster 15 (15 specimens)
Cluster 16 (295 specimens)


Cluster 17 (35 specimens)
Cluster 18 (2 specimens)
Cluster 19 (17 specimens)
Cluster 20 (87 specimens)
Cluster 21 (27 specimens)
Cluster 22 (68 specimens)


Cluster 23 (54 specimens)
Cluster 24 (17 specimens)
Cluster 25 (41 specimens)
Cluster 26 (8 specimens)
Cluster 27 (62 specimens)
Cluster 28 (37 specimens)

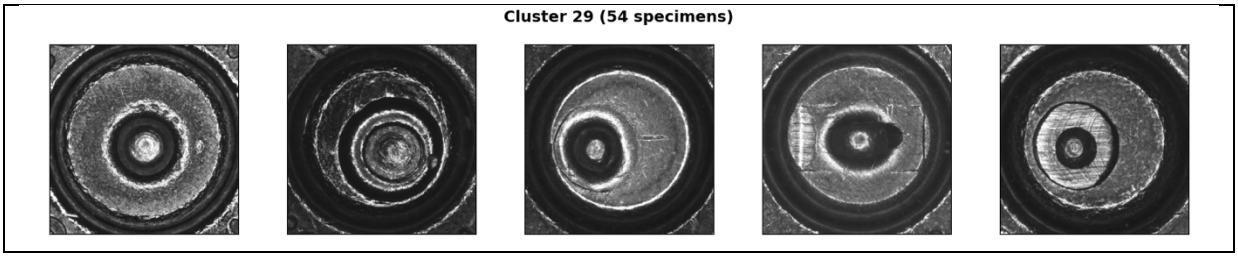


Figure-A VIII-1 Images 2D des cinq premiers échantillons de chaque cluster OPTICS, pour un nombre de clusters de 30. Extraction des caractéristiques par la CNN ENB3 et réduction à 2 dimensions par TSNE

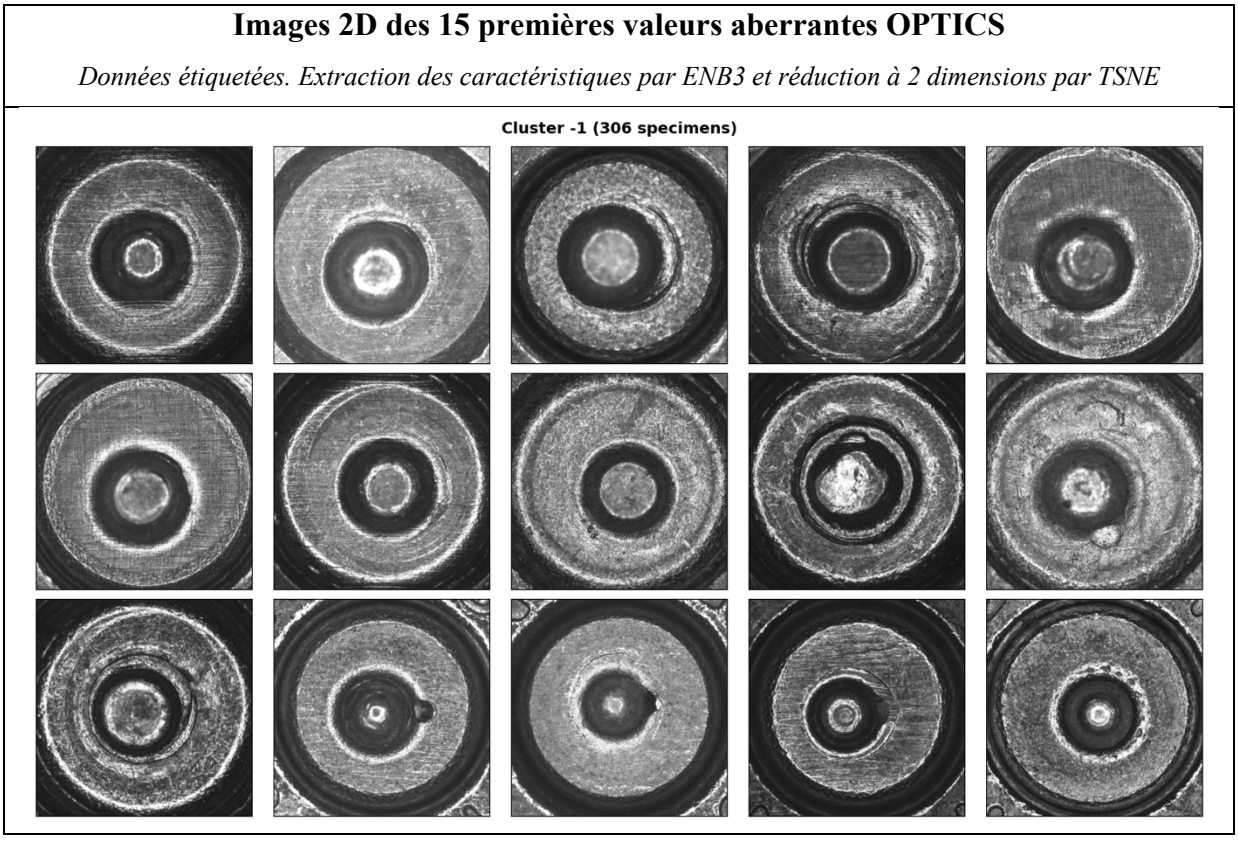


Figure-A VIII-2 Images 2D des 15 premiers échantillons considérés comme étant des valeurs aberrantes par OPTICS

ANNEXE IX

CONTRIBUTION #2 : SPECTRAL

Images 2D des clusters spectral contenant un seul échantillon

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

Paramètres : *affinity = CHI2, assign_labels = cluster_qr et n_clusters = 30*

30 clusters

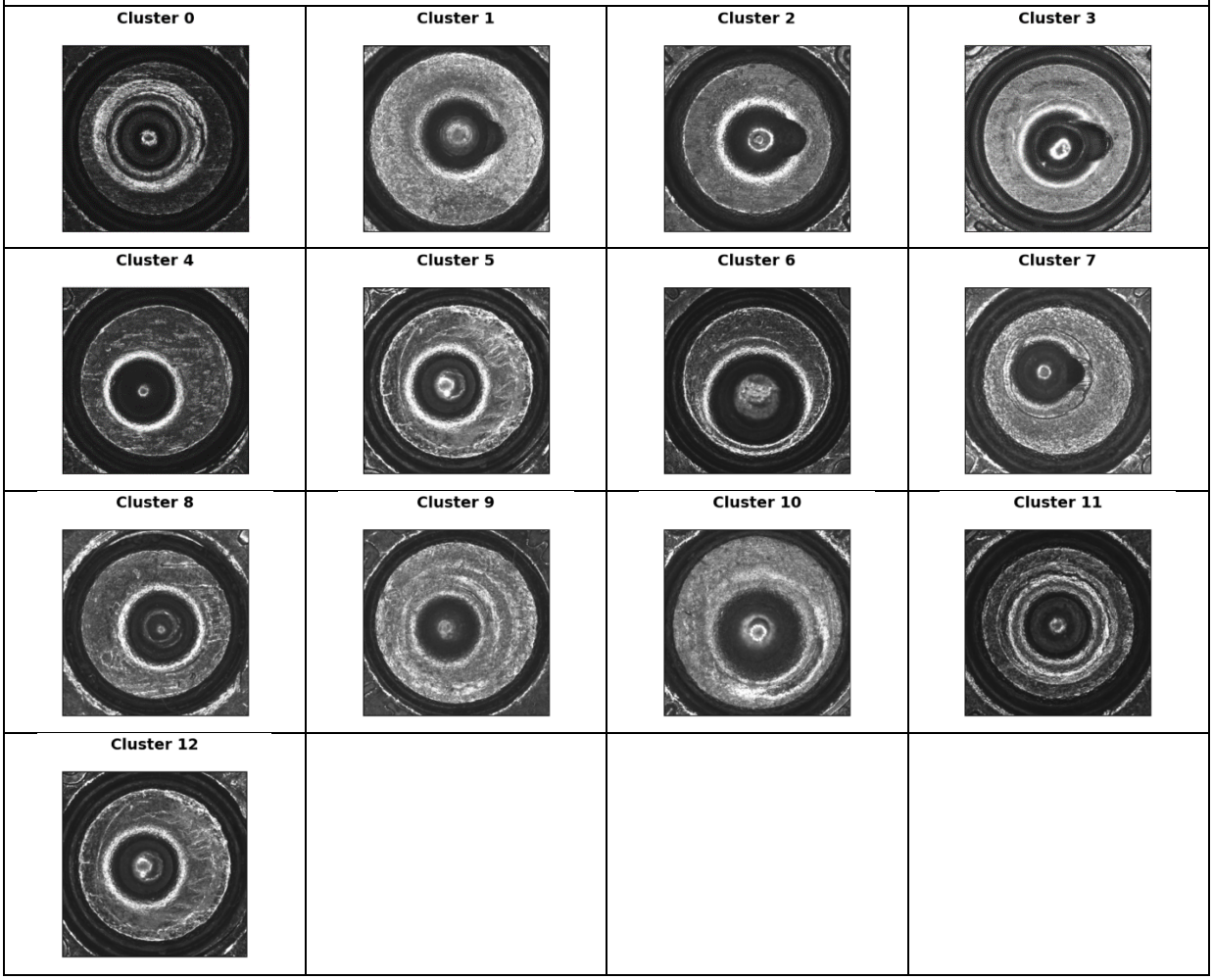


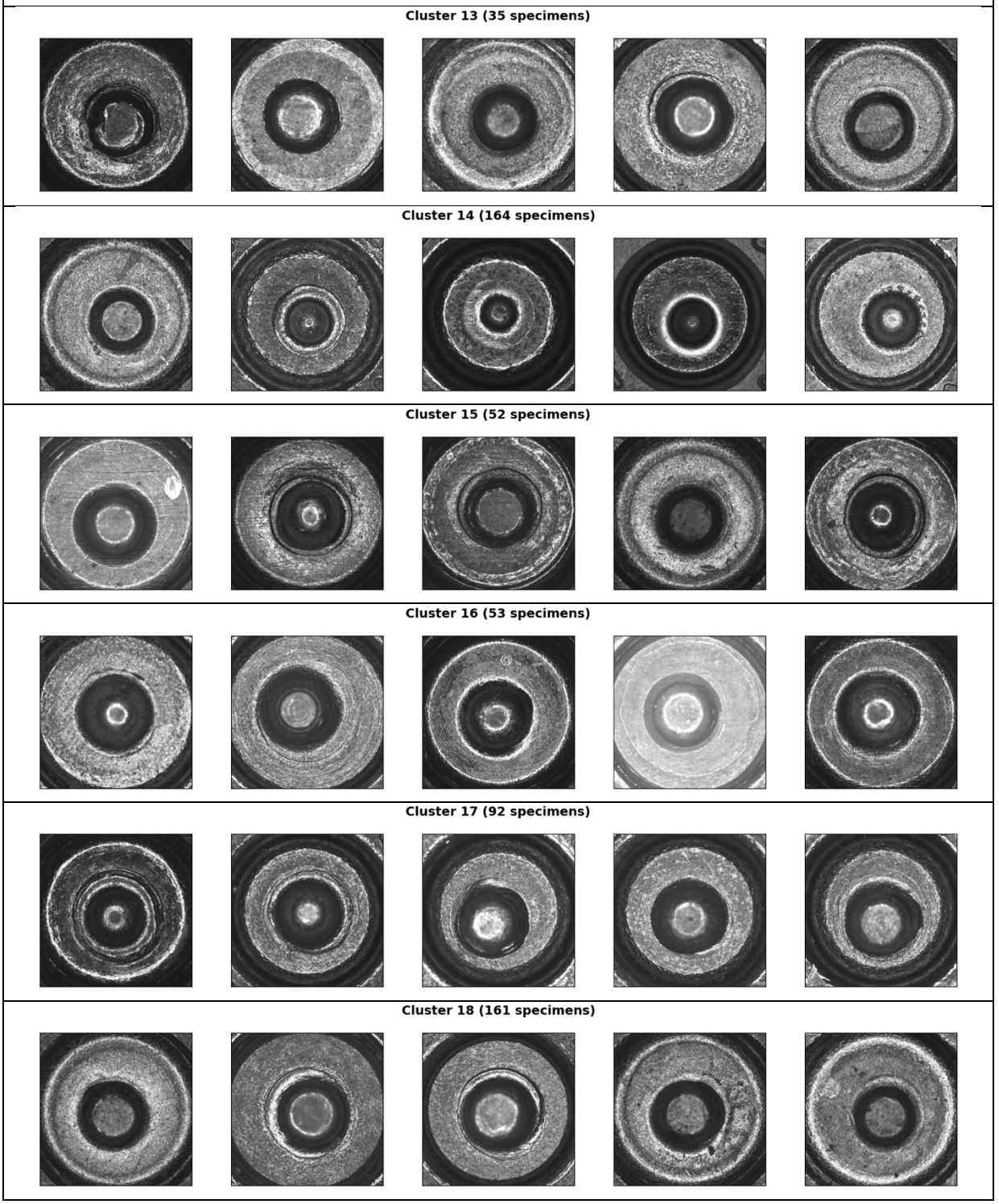
Figure-A IX-1 Images 2D des 13 cluster spectral contenant un seul échantillon, pour un nombre de clusters de 30. Extraction des caractéristiques par la CNN ENB3 et réduction à 2 dimensions par TSNE

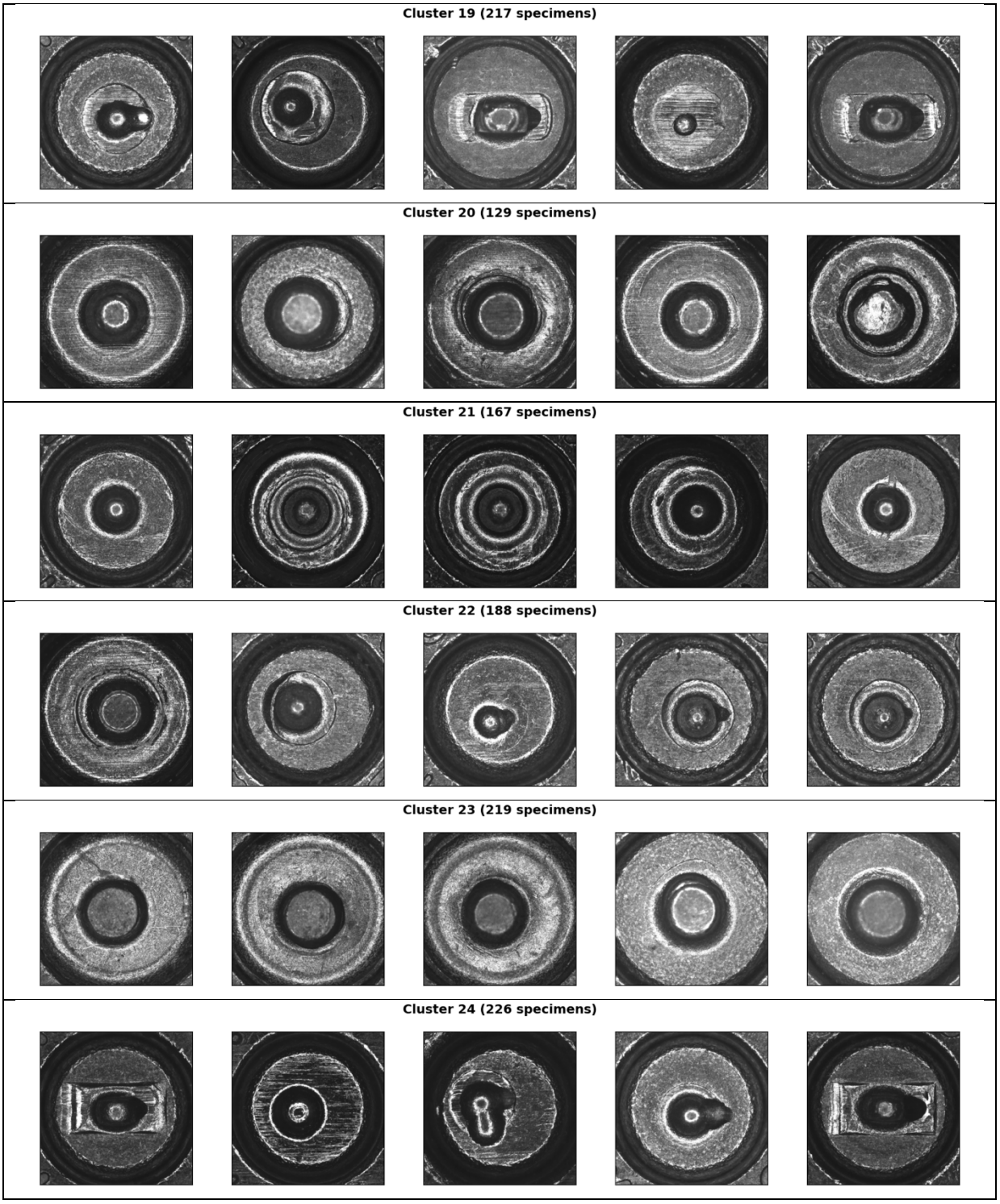
Images 2D des cinq premiers échantillons des autres cluster spectraux

Données étiquetées. Extraction des caractéristiques par ENB3 et réduction à 2 dimensions par TSNE

Paramètres : ***affinity = CHI2, assign_labels = cluster_qr et n_clusters = 30***

30 clusters




Cluster 19 (217 specimens)
Cluster 20 (129 specimens)
Cluster 21 (167 specimens)
Cluster 22 (188 specimens)
Cluster 23 (219 specimens)
Cluster 24 (226 specimens)

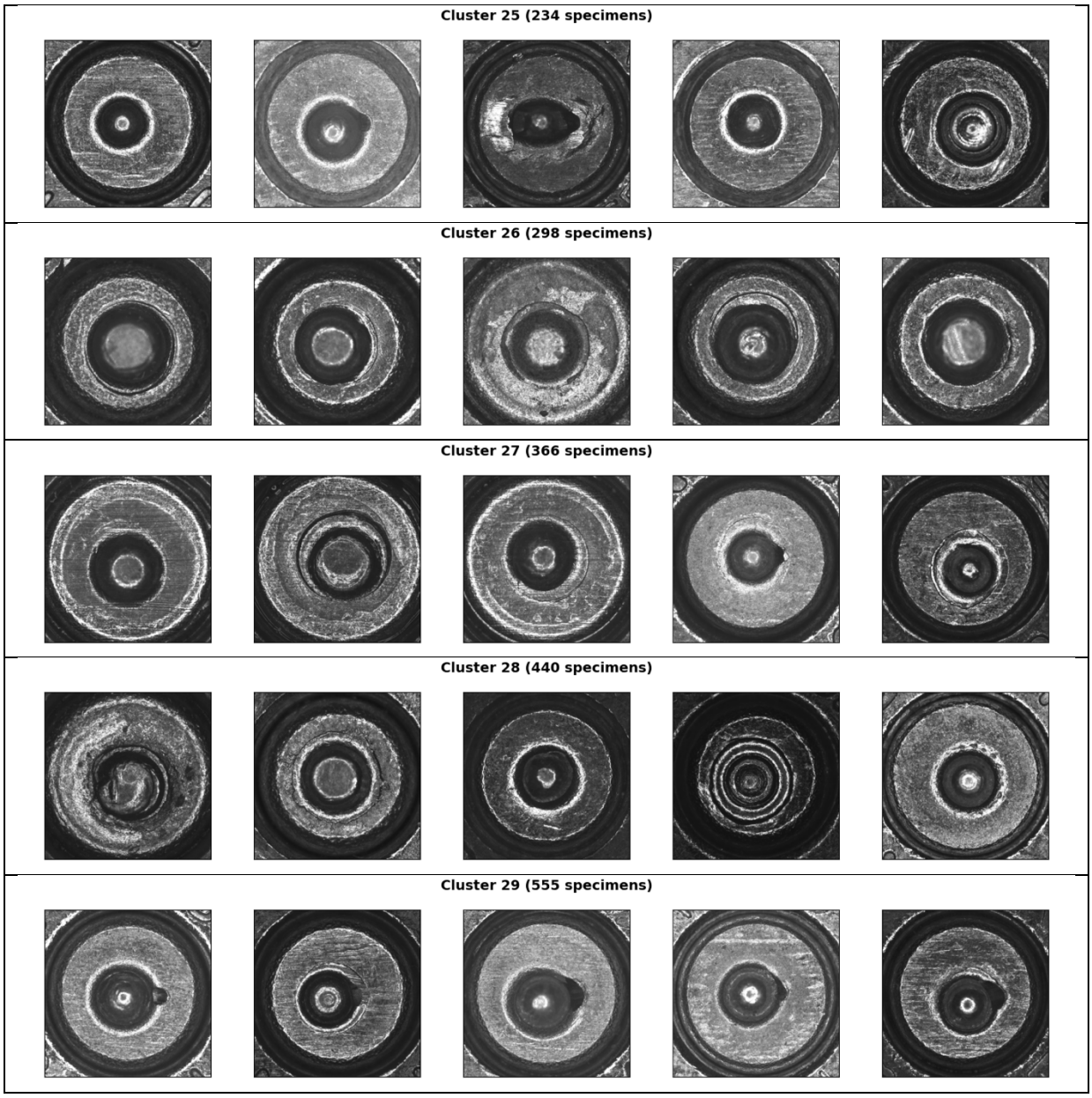


Figure-A IX-2 Images 2D des cinq premiers échantillons des 17 cluster spectraux contenant plus d'un échantillon, pour un nombre de clusters de 30. Extraction des caractéristiques par la CNN ENB3 et réduction à 2 dimensions par TSNE

ANNEXE X

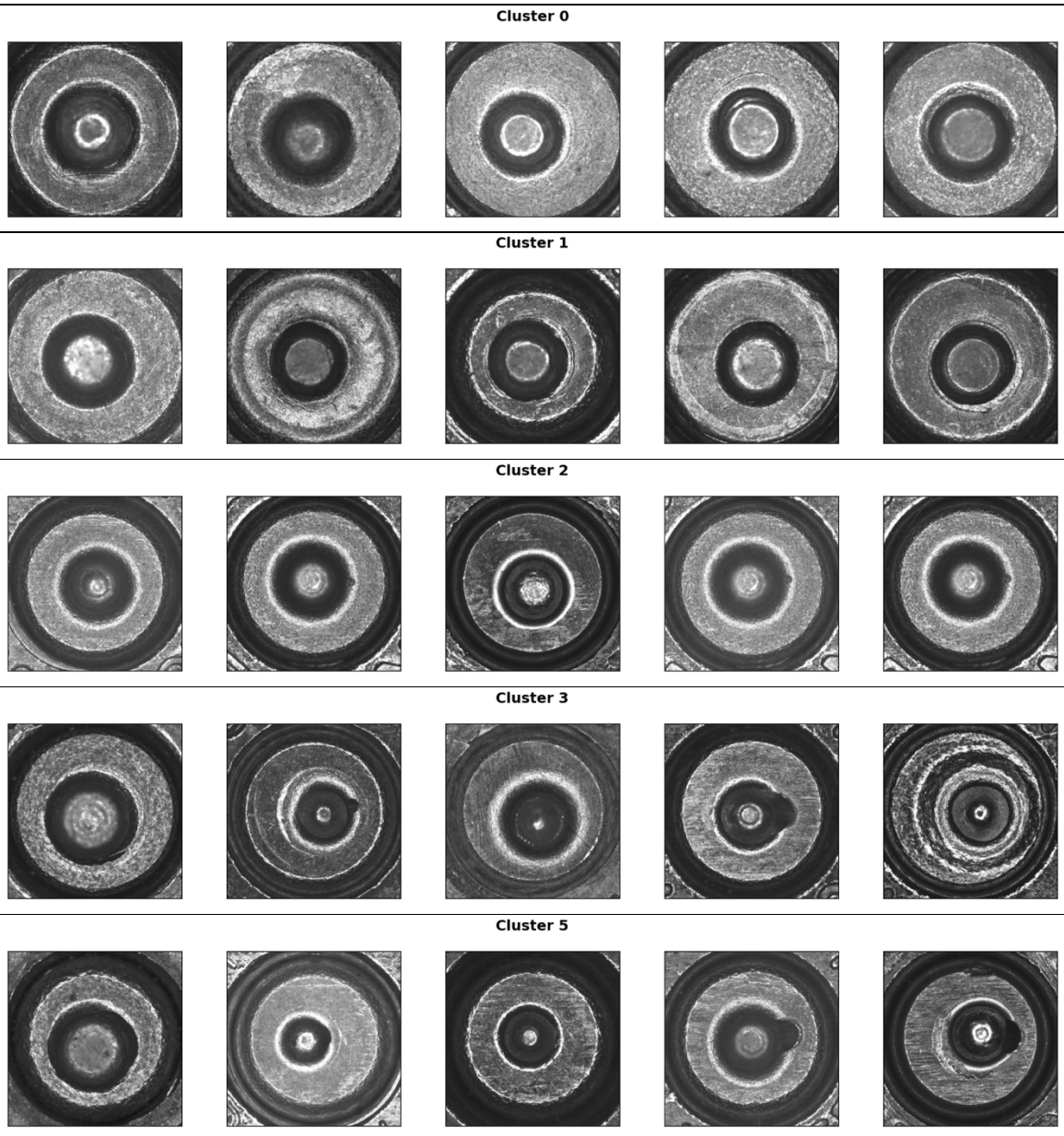
CONTRIBUTION #2 : DCN

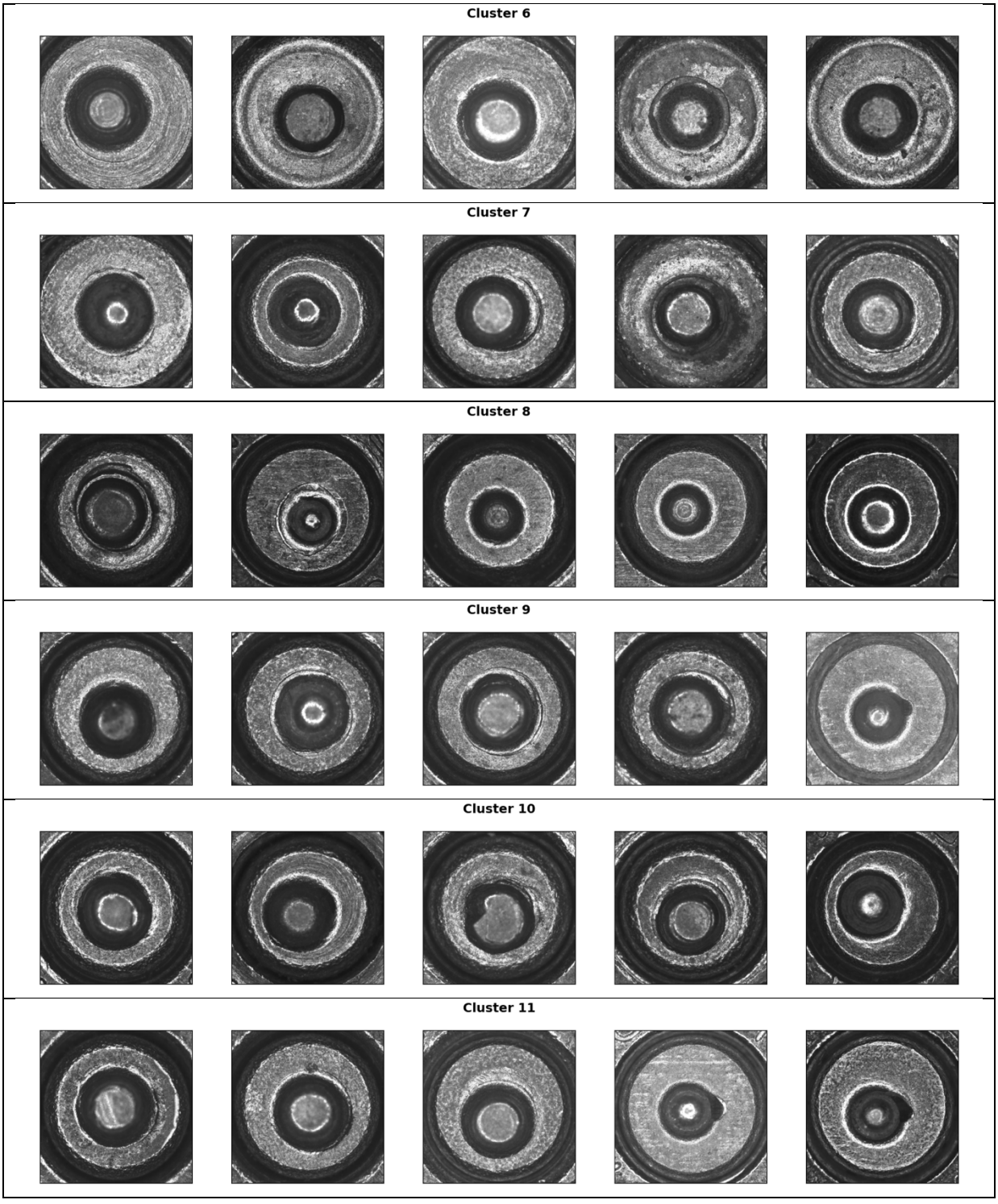
Images 2D des cinq premiers échantillons de chaque cluster DCN

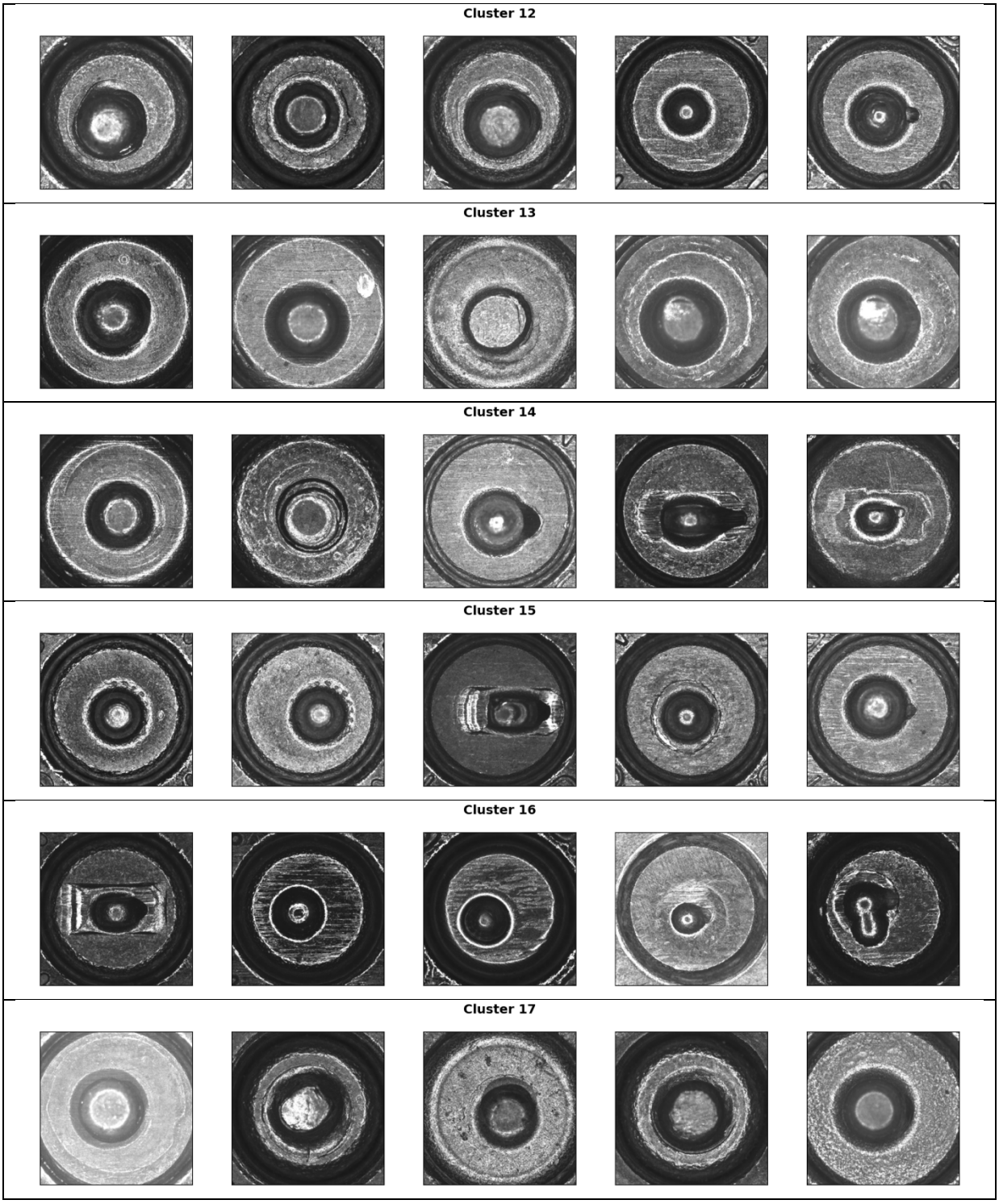
Données étiquetées avec bruit

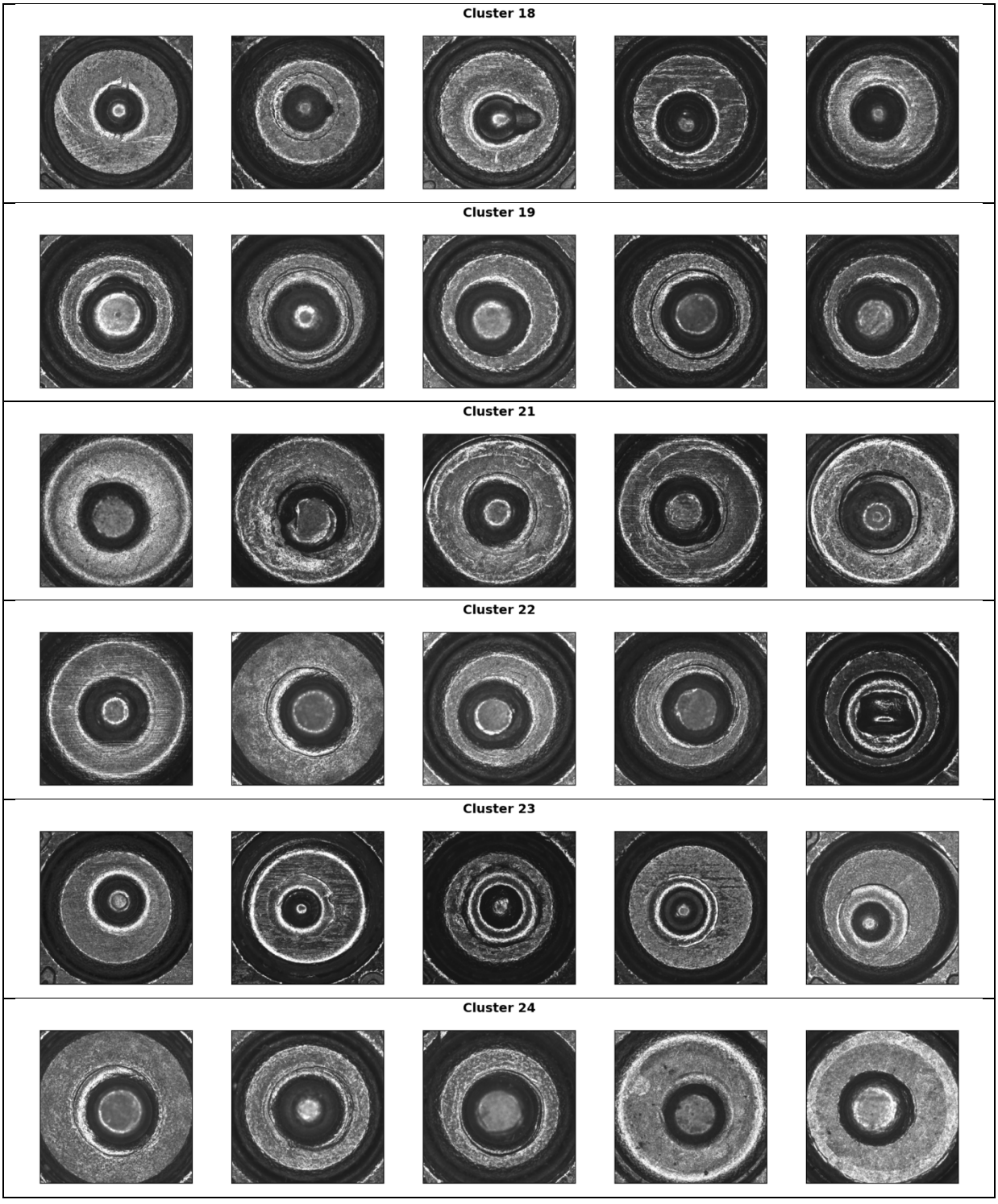
Paramètres : ***Encodeur = [2048, 1024, 256], Espace latent = 2, clust_loss_weight = 0,2***

28 clusters




Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Cluster 11


Cluster 12
Cluster 13
Cluster 14
Cluster 15
Cluster 16
Cluster 17


Cluster 18
Cluster 19
Cluster 21
Cluster 22
Cluster 23
Cluster 24

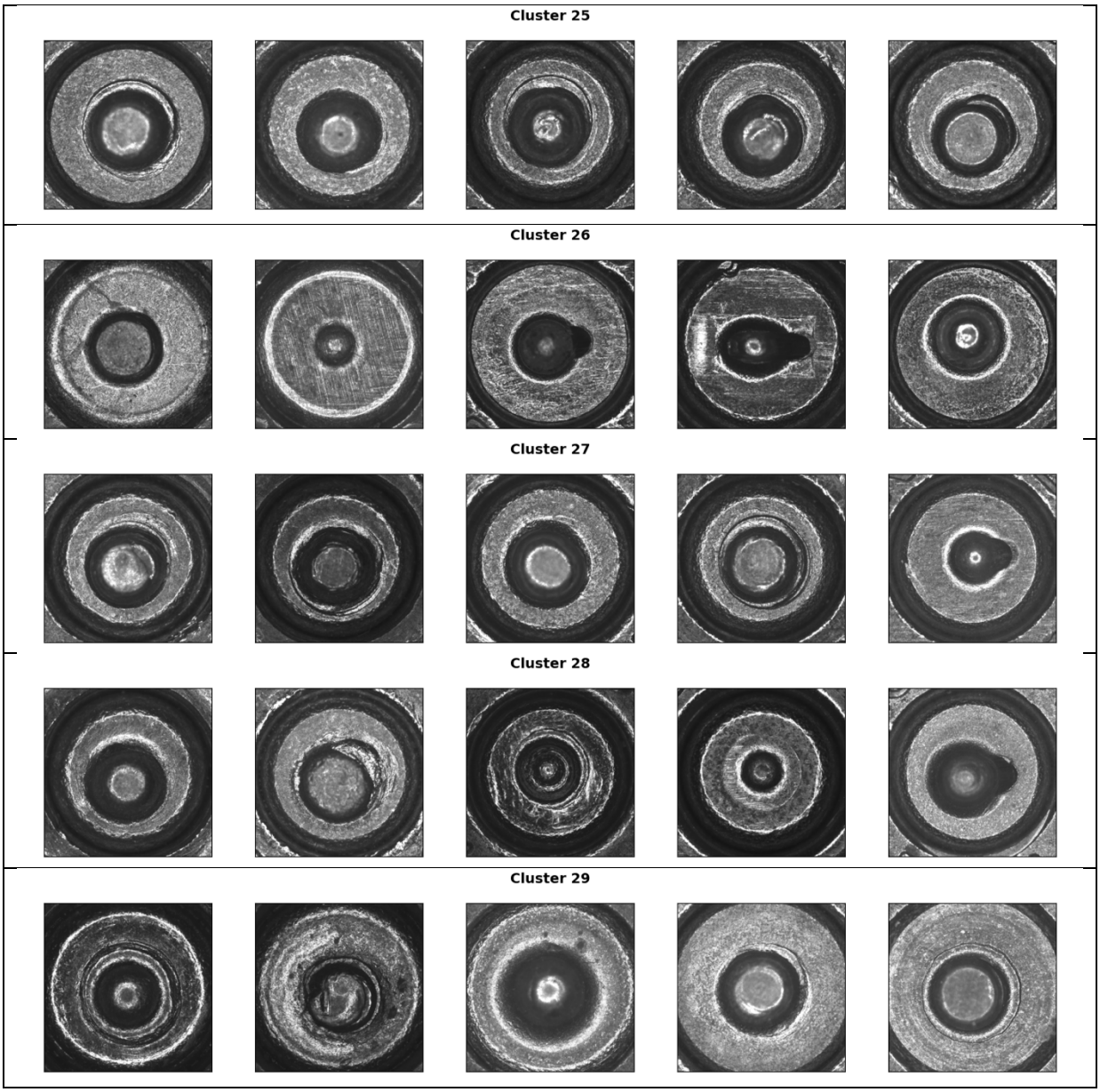


Figure-A X-1 Images 2D des cinq premiers échantillons de chaque cluster DCN, pour un nombre de clusters de 30

ANNEXE XI

CONTRIBUTION #2 : SOMMAIRE

Tableau-A XI-1 Paramètres utilisés pour les algorithmes de clustering

Algorithme	**Paramètre**	**Valeur**
K-means	*nb_clusters*	30
Fuzzy C-means	*nb_clusters*	6
	m	3
DBSCAN	*eps*	3,5
	min_samples	25
HDBSCAN	*metric*	Cosine
	min_samples	10
	min_cluster_size	95
OPTICS	*metric*	Minkowski
	min_cluster_size	5
	min_samples	15
	cluster_method	dbscan
	eps	3,0
Spectral	*affinity*	CHI2
	gamma	1,0
	assign_labels	Cluster_qr
	n_clusters	30

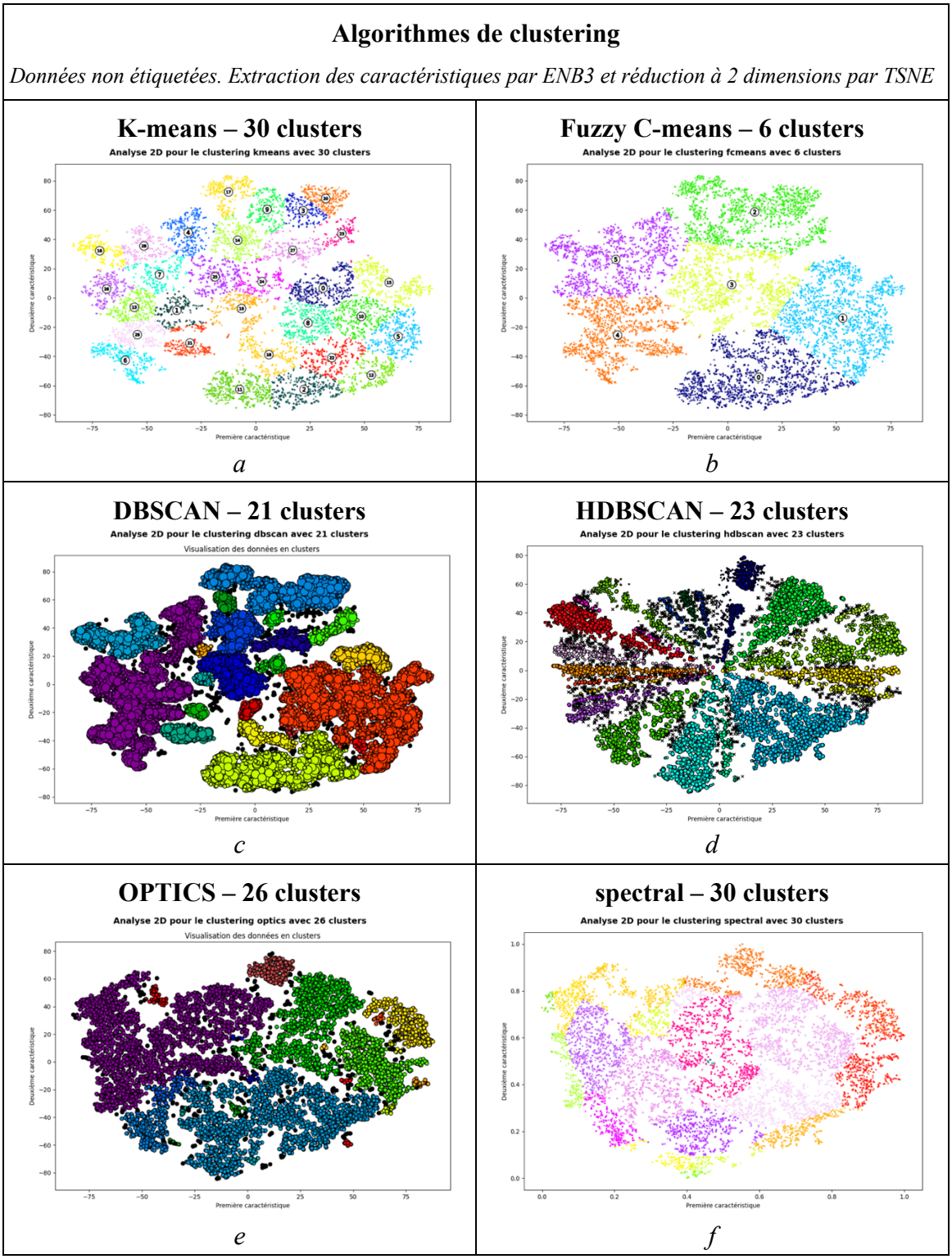


Figure-A XI-1 Graphiques en nuages de points pour les algorithmes de clustering sur les données non étiquetées, avec les paramètres sélectionnés

ANNEXE XII

CONTRIBUTION #3 : MULTIÉTIQUETTE

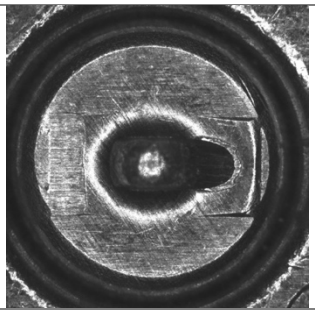
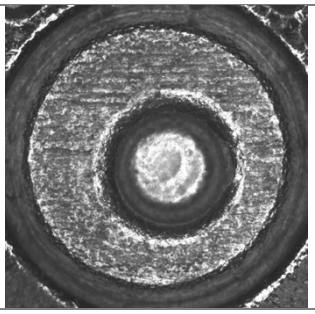
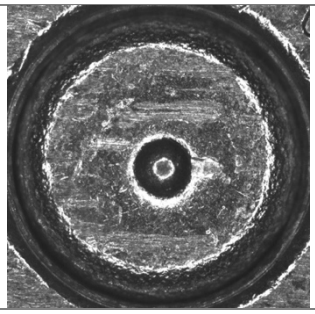
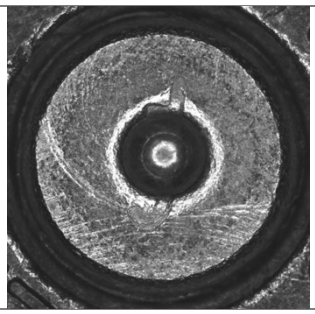
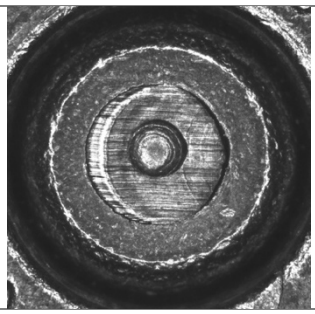
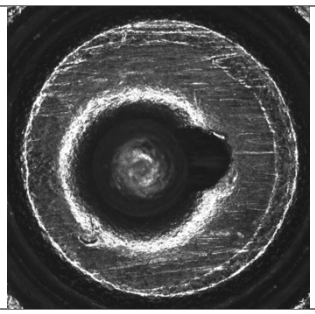
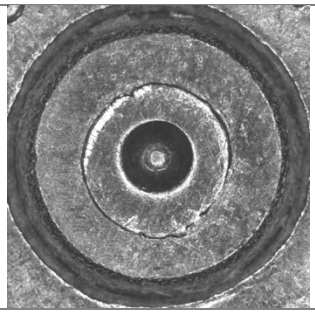
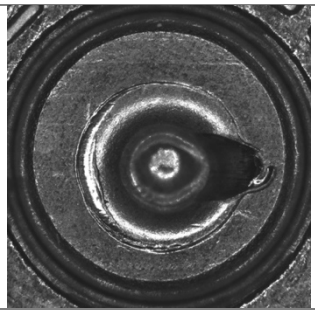
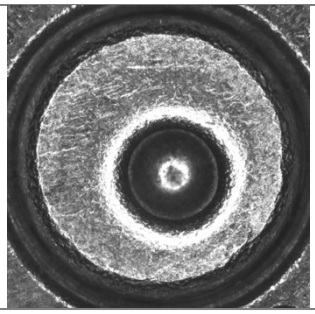
<table>
<tr><th colspan="3">Catégorie « Parallèle »</th></tr>
<tr><td colspan="3">Vrais Positifs – TP (a)</td></tr>
<tr><td>  </td><td>  </td><td>  </td></tr>
<tr><td>VT: Parallèle
Prédiction: Parallèle</td><td>VT: Parallèle, Granulaire
Prédiction: Parallèle, Granulaire</td><td>VT: Parallèle, Lisse
Prédiction: Parallèle</td></tr>
<tr><td colspan="3">Faux Positifs – FP (b)</td></tr>
<tr><td>  </td><td>  </td><td>  </td></tr>
<tr><td>VT: Arche
Prédiction: Parallèle</td><td>VT: Lisse
Prédiction: Parallèle</td><td>VT: Hachure
Prédiction: Parallèle, Hachure</td></tr>
<tr><td colspan="3">Faux Négatifs – FN (c)</td></tr>
<tr><td>  </td><td>  </td><td>  </td></tr>
<tr><td>VT: Parallèle, Lisse
Prédiction: Granulaire, Lisse</td><td>VT: Parallèle
Prédiction: Lisse</td><td>VT: Parallèle, Granulaire
Prédiction: Granulaire</td></tr>
</table>

Figure-A XII-1 Images 2D des résultats de classification pour la catégorie « Parallèle ». Modèle VT-v, avec un seuil de 65%. Vrais Positifs (a), Faux Positifs (b), Faux Négatifs (c)

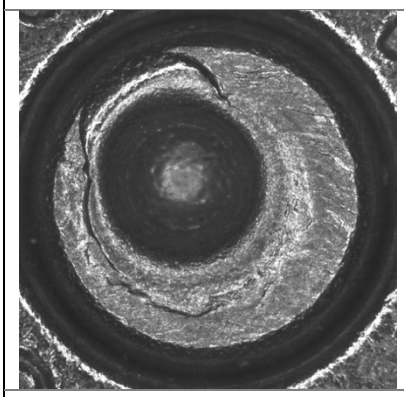
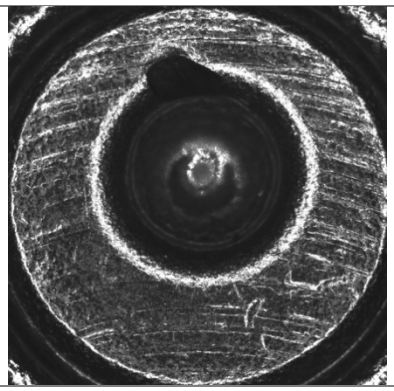
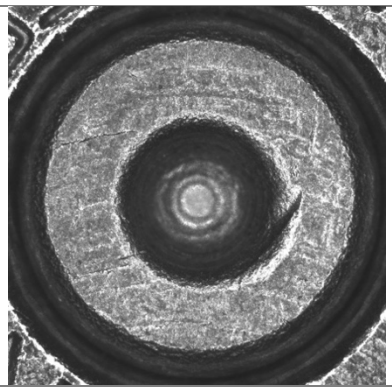
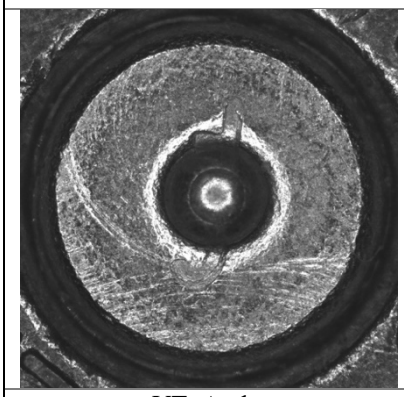
Catégorie « Arche »		
Vrais Positifs – TP (a)		
		
VT: Arche Prédiction: Parallèle, Arche	VT: Arche Prédiction: Parallèle, Arche	VT: Arche Prédiction: Parallèle, Arche
Faux Positifs – FP (b)		
Aucun		
Faux Négatifs – FN (c)		
		
VT: Arche Prédiction: Parallèle		

Figure-A XII-2 Images 2D des résultats de classification pour la catégorie « Arche ». Modèle VT-v, avec un seuil de 65%. Vrais Positifs (a), Faux Positifs (b), Faux Négatifs (c)

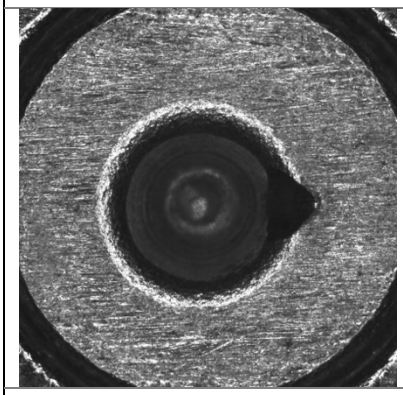
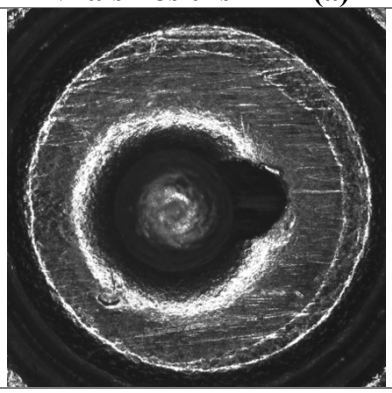
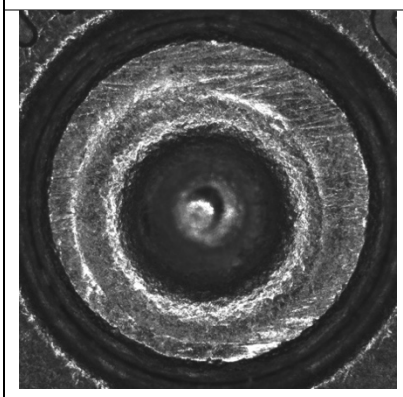
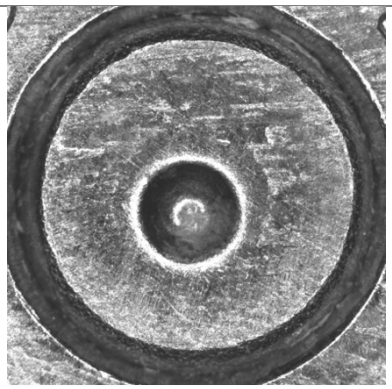
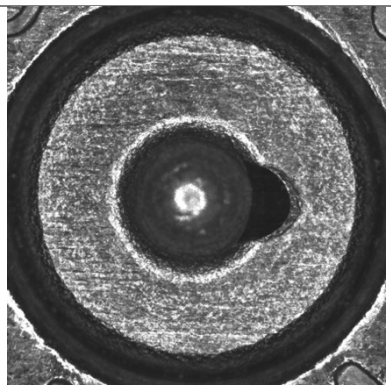
<table>
<tr><th colspan="3">Catégorie « Hachure »</th></tr>
<tr><td colspan="3">Vrais Positifs – TP (a)</td></tr>
<tr><td>  </td><td>  </td><td></td></tr>
<tr><td>GT: Parallèle, Hachure
Prédiction: Parallèle, Hachure</td><td>GT: Hachure
Prédiction: Parallèle, Hachure</td><td></td></tr>
<tr><td colspan="3">Faux Positifs – FP (b)</td></tr>
<tr><td colspan="3">Aucun</td></tr>
<tr><td colspan="3">Faux Négatifs – FN (c)</td></tr>
<tr><td>  </td><td>  </td><td>  </td></tr>
<tr><td>VT: Hachure
Prédiction: Granulaire</td><td>VT: Hachure
Prédiction: Parallèle</td><td>VT: Hachure
Prédiction: Parallèle</td></tr>
</table>

Figure-A XII-3 Images 2D des résultats de classification pour la catégorie « Hachure ». Modèle VT-v, avec un seuil de 65%. Vrais Positifs (a), Faux Positifs (b), Faux Négatifs (c)

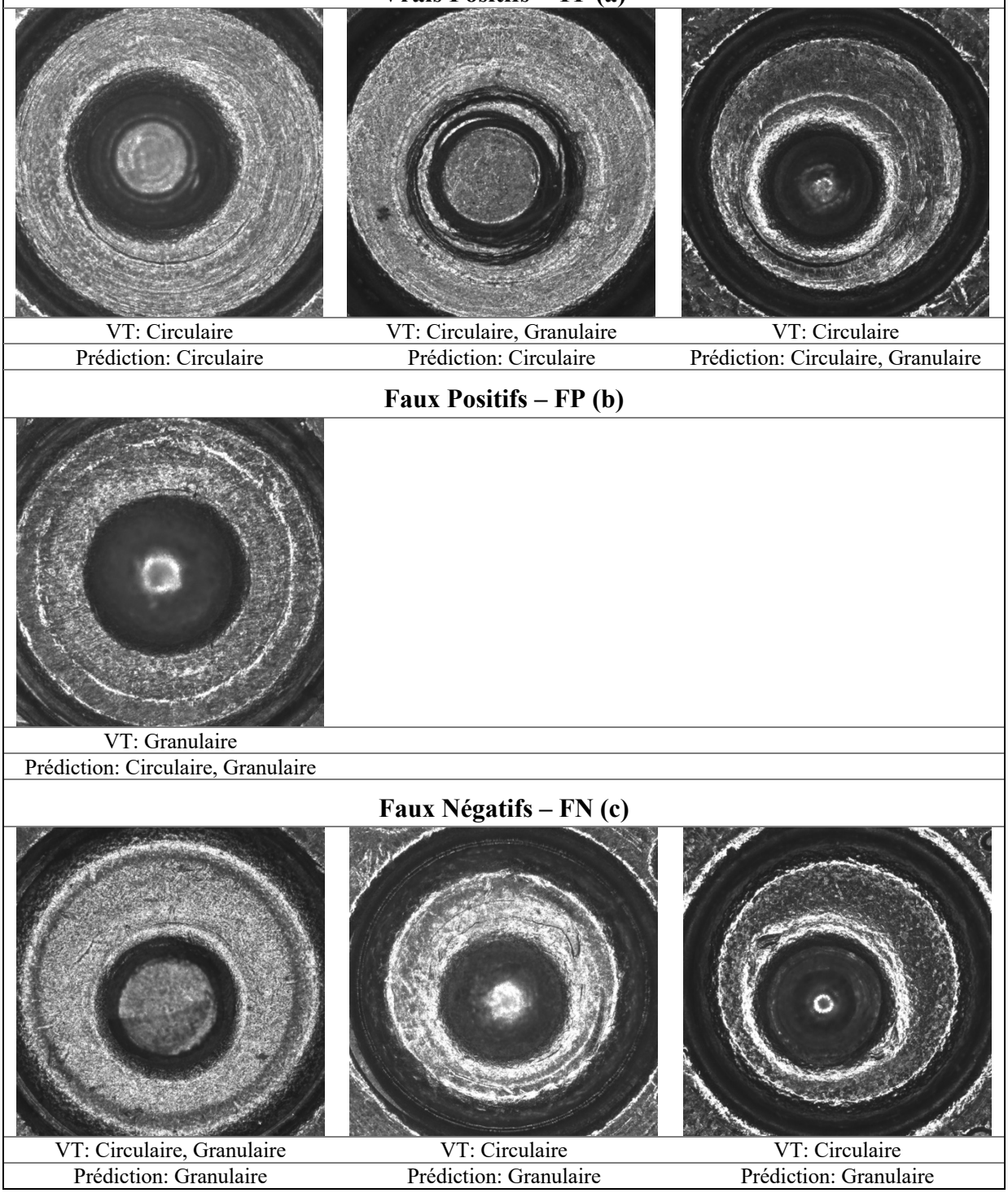


Figure-A XII-4 Images 2D des résultats de classification pour la catégorie « Circulaire ». Modèle VT-v, avec un seuil de 65%. Vrais Positifs (a), Faux Positifs (b), Faux Négatifs (c)

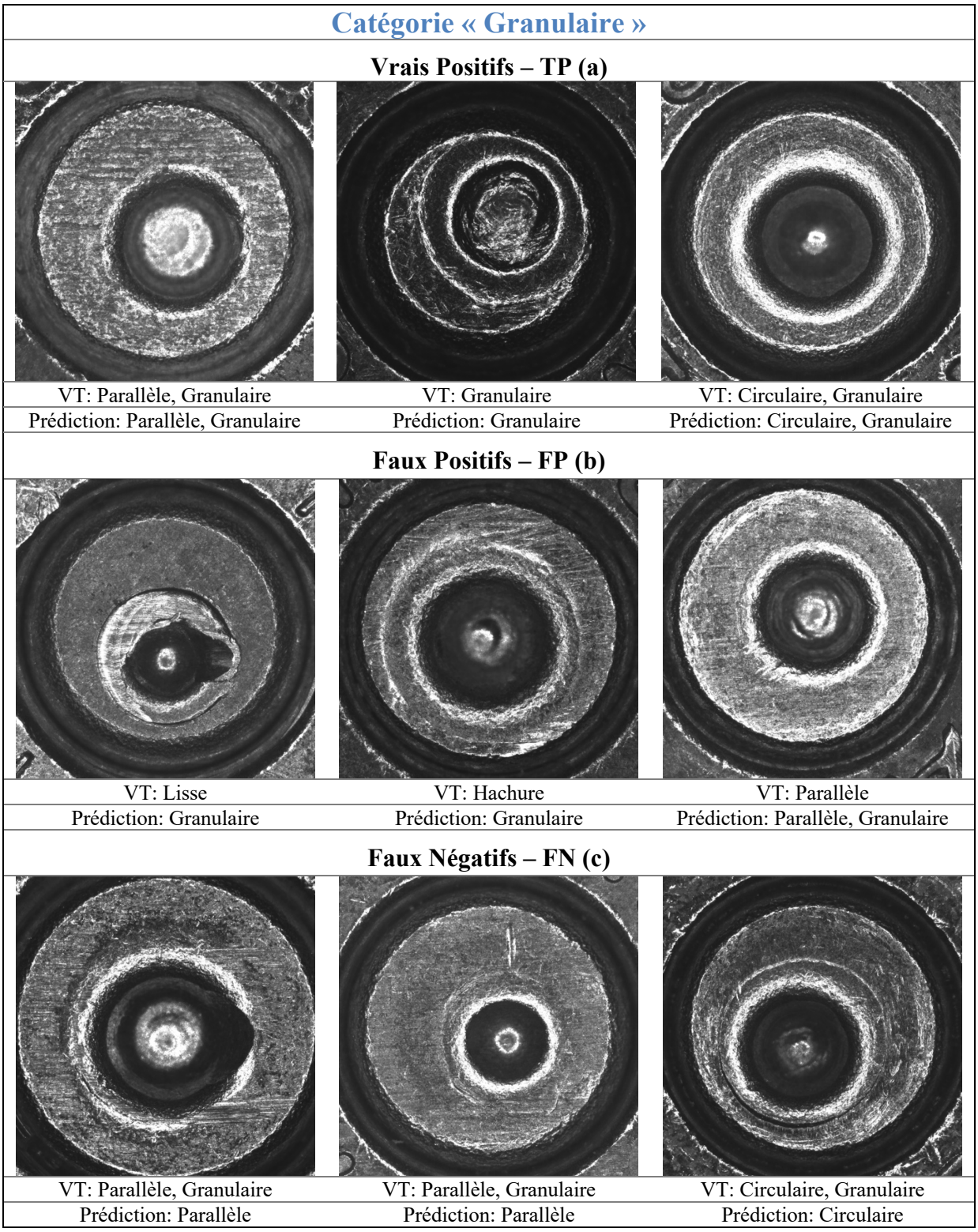


Figure-A XII-5 Images 2D des résultats de classification pour la catégorie « Granulaire ». Modèle VT-v, avec un seuil de 65%. Vrais Positifs (a), Faux Positifs (b), Faux Négatifs (c)

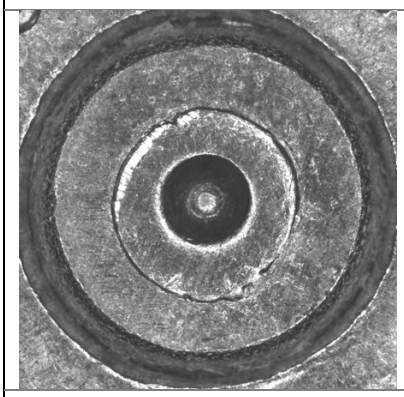
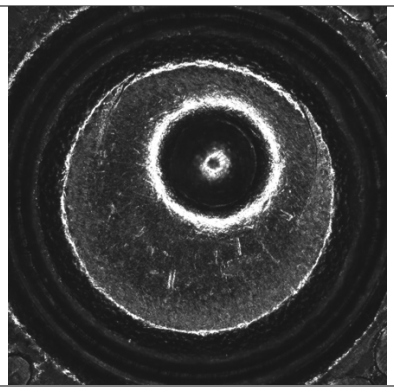
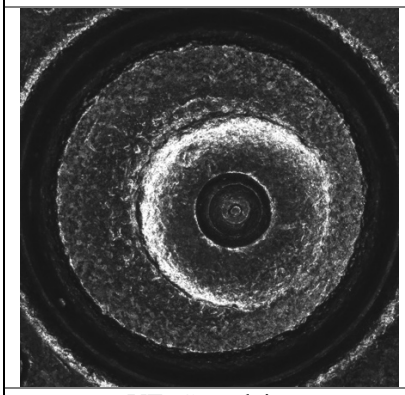
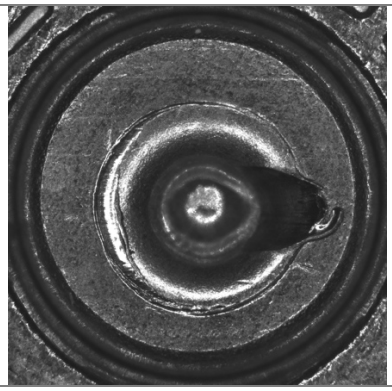
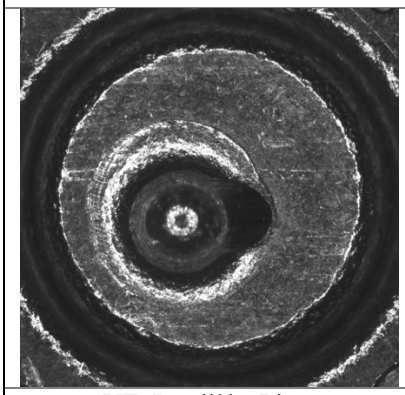
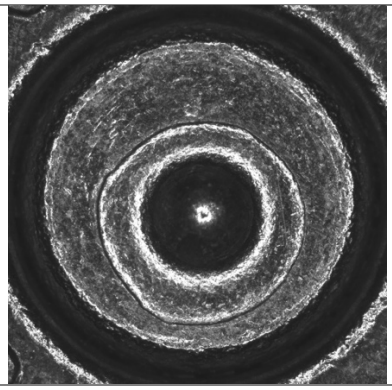
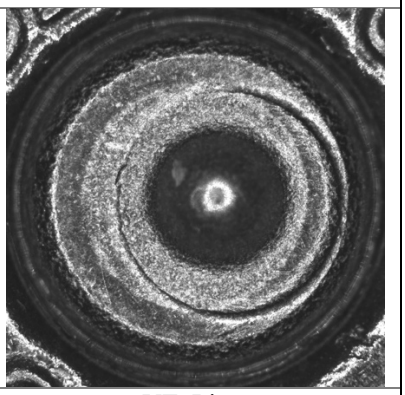
<table>
<tr><th colspan="3">Catégorie « Lisse »</th></tr>
<tr><td colspan="3">Vrais Positifs – TP (a)</td></tr>
<tr><td>  </td><td>  </td><td></td></tr>
<tr><td>VT: Parallèle, Lisse
Prédiction: Granulaire, Lisse</td><td>VT: Lisse
Prédiction: Granulaire, Lisse</td><td></td></tr>
<tr><td colspan="3">Faux Positifs – FP (b)</td></tr>
<tr><td>  </td><td>  </td><td></td></tr>
<tr><td>VT: Granulaire
Prédiction: Granulaire, Lisse</td><td>VT: Parallèle
Prédiction: Lisse</td><td></td></tr>
<tr><td colspan="3">Faux Négatifs – FN (c)</td></tr>
<tr><td>  </td><td>  </td><td>  </td></tr>
<tr><td>VT: Parallèle, Lisse
Prédiction: Parallèle</td><td>VT: Circulaire, Lisse
Prédiction: Circulaire</td><td>VT: Lisse
Prédiction: Granulaire</td></tr>
</table>

Figure-A XII-6 Images 2D des résultats de classification pour la catégorie « Lisse ». Modèle VT-v, avec un seuil de 65%. Vrais Positifs (a), Faux Positifs (b), Faux Négatifs (c)

LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES

AFTE. 2013. *AFTE Glossary - Glossary of the Association of Firearm and Tool Mark Examiners, 6th edition*. <https://afte.org/uploads/documents/AFTE_Glossary_Version_6.110619_DRAFT_.PDF>. Consulté le 30 mai 2022.

Amnesty International. [s d]. « Amnesty International ». In *Amnesty International*. <https://www.amnesty.org/en/>. Consulté le 25 mars 2025.

Atila, Ümit, Murat Uçar, Kemal Akyol et Emine Uçar. 2021. « Plant leaf disease classification using EfficientNet deep learning model ». *Ecological Informatics*, vol. 61, p. 101182. <https://doi.org/10.1016/j.ecoinf.2020.101182>.

Byeon, Wonmin, Manuel Domínguez-Rodrigo, Georgios Arampatzis, Enrique Baquedano, José Yravedra, Miguel Angel Maté-González et Petros Koumoutsakos. 2019. « Automated identification and deep classification of cut marks on bones and its paleoanthropological implications ». *Journal of Computational Science*, vol. 32, p. 36-43. <https://doi.org/10.1016/j.jocs.2019.02.005>.

Cai, Yunliang et George Baciu. 2013. « Detecting, Grouping, and Structure Inference for Invariant Repetitive Patterns in Images ». *IEEE Transactions on Image Processing*, vol. 22, nº 6, p. 2343-2355. <https://doi.org/10.1109/TIP.2013.2251649>.

Dahal, Parmas. 2019. « Deep Clustering: Using Deep Neural Networks for Clustering ». <https://www.parasdahal.com/deep-clustering>. Consulté le 18 juillet 2023.

Dias, Luiz Dantas. [s d]. *fuzzy-c-means: A simple python implementation of Fuzzy C-means algorithm.* <https://github.com/omadson/fuzzy-c-means>. Consulté le 24 février 2025.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit et Neil Houlsby. 2021. « An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale ». In *ICLR 2021*. (3 juin 2021). arXiv. <http://arxiv.org/abs/2010.11929>. Consulté le 18 mai 2022.

Education Ecosystem, LEDU. 2018. « Understanding K-means Clustering in Machine Learning ». *Towards Data Science*. <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>.

Fischer, B. et J.M. Buhmann. 2003. « Path-based clustering for grouping of smooth curves and texture segmentation ». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, nº 4, p. 513-518. <https://doi.org/10.1109/TPAMI.2003.1190577>.

Fischer, Robert et Claus Vielhauer. 2015. « Towards automated firearm identification based on high resolution 3D data: rotation-invariant features for multiple line-profile-measurement of firing pin shapes ». In *IS&T/SPIE Electronic Imaging*. (San Francisco, California, United States, 17 mars 2015), p. 93930Q. Proc SPIE Int Soc Opt Eng. <https://doi.org/10.1117/12.2077567>.

Gambino, Carol, Patrick McLaughlin, Loretta Kuo, Frani Kammerman, Peter Shenkin, Peter Diaczuk, Nicholas Petraco, James Hamby et Nicholas D. K. Petraco. 2011. « Forensic surface metrology: tool mark evidence ». *Scanning*, vol. 33, nº 5, p. 272-278. <https://doi.org/10.1002/sca.20251>.

Gerules, George, Sanjiv K. Bhatia et Daniel E. Jackson. 2013. « A survey of image processing techniques and statistics for ballistic specimens in forensic science ». *Science & Justice*, vol. 53, nº 2, p. 236-250. <https://doi.org/10.1016/j.scijus.2012.07.002>.

Glrx. 2021. *Structure d'une cartouche*. <https://commons.wikimedia.org/wiki/File:Cartridge_cross_section.svg>. Consulté le 17 juillet 2025.

Goodfellow, Ian, Yoshua Bengio et Aaron Courville. 2016. *Deep learning*. Coll. « Adaptive computation and machine learning ». Cambridge, Massachusetts : The MIT Press, 775 p.

Kaddioui, Houda, Luc Duong, Julie Joncas, Christian Bellefleur, Imad Nahle, Olivier Chémaly, Marie-Lyne Nault, Stefan Parent, Guy Grimard et Hubert Labelle. 2020. « Convolutional Neural Networks for Automatic Risser Stage Assessment ». *Radiology: Artificial Intelligence*, vol. 2, nº 3, p. e180063. <https://doi.org/10.1148/ryai.2020180063>.

Kara, Ilker. 2016. « Investigation of Ballistic Evidence through an Automatic Image Analysis and Identification System ». *Journal of Forensic Sciences*, vol. 61, nº 3, p. 775-781. <https://doi.org/10.1111/1556-4029.13073>.

Kara, Ilker et Alihan Karatatar. 2022. « Classification of fired cartridge cases using 3D image capture and a comparison of database correlation method performance ». *Journal of Forensic Sciences*, vol. 67, nº 5, p. 1998-2008. <https://doi.org/10.1111/1556-4029.15089>.

Kaur, Navneet, Nahida Nazir, et Manik. 2021. « A Review of Local Binary Pattern Based texture feature extraction ». In *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. (septembre 2021), p. 1-4. <https://doi.org/10.1109/ICRITO51393.2021.9596485>.

Keras. [s d]. « Keras documentation: VGG16 and VGG19 ». <https://keras.io/api/applications/vgg/>. Consulté le 14 décembre 2022a.

Keras. [s d]. « Keras documentation: EfficientNet B0 to B7 ». <https://keras.io/api/applications/efficientnet/>. Consulté le 14 août 2024b.

Kubota, Yasuhiro. [s d]. « tf-keras-vis v0.8.7 documentation ». <https://keisen.github.io/tf-keras-vis-docs/>. Consulté le 20 août 2024.

Kudonu, Moza, Mohammad A. AlShamsi, Sharon Philip, Gursirat Khokhar, Parli B. Hari et Nrashant Singh. 2022. « Artificial Intelligence: Future of Firearm Examination ». In *2022 Advances in Science and Engineering Technology International Conferences (ASET)*. (Dubai, United Arab Emirates, 21 février 2022), p. 1-5. IEEE. <https://doi.org/10.1109/ASET53988.2022.9735105>.

Kundu, B., K.P. White et C. Mastrangelo. 2002. « Defect clustering and classification for semiconductor devices ». In *The 2002 45th Midwest Symposium on Circuits and Systems, 2002. MWSCAS-2002.* (août 2002), p. II-II. <https://doi.org/10.1109/MWSCAS.2002.1186923>.

Le Bouthillier, Marie-Eve, Lynne Hrynkiw, Alain Beauchamp, Luc Duong et Sylvie Ratté. 2023. « Automated detection of regions of interest in cartridge case images using deep learning ». *Journal of Forensic Sciences*, vol. n/a, nº n/a. <https://doi.org/10.1111/1556-4029.15319>. Consulté le 11 septembre 2023.

Li, Katherine Shu-Min, Leon Li-Yang Chen, Ken Chau-Cheung Cheng, Peter Yi-Yu Liao, Sying-Jyan Wang, Andrew Yi-An Huang, Nova Tsai, Leon Chou, Gus Chang-Hung Han, Jwu E Chen, Hsing-Chung Liang et Chun-Lung Hsu. 2021. « Automatic Inspection for Wafer Defect Pattern Recognition with Unsupervised Clustering ». In *2021 IEEE European Test Symposium (ETS)*. (mai 2021), p. 1-2. <https://doi.org/10.1109/ETS50041.2021.9465457>.

LO. [s d]. « LeadsOnline: Empowering Law Enforcement with Data & Forensic Solutions ». <https://leadsonline.com>. Consulté le 16 juin 2025.

Manimozhi, I. et S. Janakiraman. 2019. « Defect detection in pattern texture analysis using improved support vector machine ». *Cluster Computing*, vol. 22, nº S6, p. 15223-15230. <https://doi.org/10.1007/s10586-018-2551-y>.

Morales, Fausto. [s d]. *vit-keras: Keras implementation of ViT (Vision Transformer)*. <https://github.com/faustomorales/vit-keras>. Consulté le 15 août 2024.

Morris. 2021. *stallmo/dcn-deep-clustering-networks*. <https://github.com/stallmo/dcn-deep-clustering-networks>. Consulté le 14 avril 2025.

Morris, Keith B., Eric F. Law, Roger L. Jefferys, Elizabeth C. Dearth et Emily B. Fabyanic. 2017. « An evaluation of the discriminating power of an Integrated Ballistics Identification System ® Heritage™ system with the NIST standard cartridge case

(Standard Reference Material 2461) ». *Forensic Science International*, vol. 280, p. 188-193. <https://doi.org/10.1016/j.forsciint.2017.09.004>.

Ngan, Henry Y. T., Grantham K. H. Pang et Nelson H. C. Yung. 2010. « Performance Evaluation for Motif-Based Patterned Texture Defect Detection ». *IEEE Transactions on Automation Science and Engineering*, vol. 7, nº 1, p. 58-72. <https://doi.org/10.1109/TASE.2008.2005418>.

Ngan, Henry Y.T., Grantham K.H. Pang et Nelson H.C. Yung. 2007. « Patterned Fabric Defect Detection using a Motif-Based Approach ». In *2007 IEEE International Conference on Image Processing*. (septembre 2007), p. II-33- II-36. <https://doi.org/10.1109/ICIP.2007.4379085>.

NIBIN. [s d]. « National Integrated Ballistic Information Network ». In *ATF Bureau of Alcohol, Tobacco, Firearms and Explosives*. <https://www.atf.gov/firearms/national-integrated-ballistic-information-network-nibin>. Consulté le 20 juin 2022.

NIST. 2012. « Cartridge Case ». *NIST*. <https://www.nist.gov/image/cartridgecasejpg>. Consulté le 17 juillet 2025.

NIST. 2015. « New Ballistics Control Chart for Forensic Imaging ». *NIST*. <https://doi.org/10/new-ballistics-control-chart-forensic-imaging>. Consulté le 17 juillet 2025.

Perkonigg, Matthias, Daniel Sobotka, Ahmed Ba-Ssalamah et Georg Langs. 2020. *Unsupervised deep clustering for predictive texture pattern discovery in medical images*. <http://arxiv.org/abs/2002.03721>. Consulté le 1 octobre 2022.

Pisantanaroj, Pattranit, Pimlapus Tanpisuth, Piyawut Sinchavanwat, Siriporn Phasuk, Phongphan Phienphanich, Parinton Jangtawee, Kittisak Yakoompai, Montri Donphoongpi, Sanong Ekgasit et Charturong Tantibundhit. 2020. « Automated Firearm Classification From Bullet Markings Using Deep Learning ». *IEEE Access*, vol. 8, p. 78236-78251. <https://doi.org/10.1109/ACCESS.2020.2989673>.

Riva, Fabiano et Christophe Champod. 2014. « Automatic Comparison and Evaluation of Impressions Left by a Firearm on Fired Cartridge Cases ». *Journal of Forensic Sciences*, vol. 59, nº 3, p. 637-647. <https://doi.org/10.1111/1556-4029.12382>.

Riva, Fabiano, Rob Hermsen, Erwin Mattijssen, Pascal Pieper et Christophe Champod. 2017. « Objective Evaluation of Subclass Characteristics on Breech Face Marks ». *Journal of Forensic Sciences*, vol. 62, nº 2, p. 417-422. <https://doi.org/10.1111/1556-4029.13274>.

Riva, Fabiano, Erwin J. A. T. Mattijssen, Rob Hermsen, Pascal Pieper, W. Kerkhoff et Christophe Champod. 2020. « Comparison and interpretation of impressed marks left by a firearm on cartridge cases – Towards an operational implementation of a

likelihood ratio based technique ». *Forensic Science International*, vol. 313, p. 110363. <https://doi.org/10.1016/j.forsciint.2020.110363>.

Scikit-learn. [s d]. « scikit-learn: machine learning in Python — scikit-learn 1.3.0 documentation ». <https://scikit-learn.org/stable/index.html>. Consulté le 20 juillet 2023a.

Scikit-learn. [s d]. « sklearn.metrics ». In *scikit-learn*. <https://scikit-learn/stable/api/sklearn.metrics.html>. Consulté le 14 août 2024b.

Scikit-learn. [s d]. « sklearn.decomposition.PCA ». In *scikit-learn*. <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>. Consulté le 14 décembre 2022c.

Scikit-learn. [s d]. « sklearn.manifold.TSNE ». In *scikit-learn*. <https://scikit-learn/stable/modules/generated/sklearn.manifold.TSNE.html>. Consulté le 27 février 2025d.

Scikit-learn. [s d]. « Scikit-Learn: Clustering ». In *scikit-learn*. <https://scikit-learn.org/stable/modules/clustering.html>. Consulté le 18 juillet 2023e.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh et Dhruv Batra. 2020. « Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization ». *International Journal of Computer Vision*, vol. 128, nº 2, p. 336-359. <https://doi.org/10.1007/s11263-019-01228-7>.

Simonyan, Karen et Andrew Zisserman. 2015. « Very deep convolutional networks for large-scale image recognition ». In *3rd International Conference on Learning Representations, ICLR 2015, May 7, 2015 - May 9, 2015*. (San Diego, CA, United states, 2015). International Conference on Learning Representations, ICLR.

Smith, Clifton L. 2001. « Multi-dimensional cluster analysis of class characteristics for ballistics specimen identification ». In *IEEE Annual International Carnahan Conference on Security Technology, Proceedings*. (2001), p. 115-121.

Song, John, Wei Chu, Mingsi Tong et Johannes Soons. 2014. « 3D topography measurements on correlation cells—a new approach to forensic ballistics identifications ». *Measurement Science and Technology*, vol. 25, nº 6, p. 064005. <https://doi.org/10.1088/0957-0233/25/6/064005>.

Tai, Xiao Hui et William F. Eddy. 2018. « A Fully Automatic Method for Comparing Cartridge Case Images ». *Journal of Forensic Sciences*, vol. 63, nº 2, p. 440-448. <https://doi.org/10.1111/1556-4029.13577>.

Tamasflex. 2010. *Comparison microscope*. <https://commons.wikimedia.org/wiki/File:ComparisonMicroscope.png>. Consulté le 16 juin 2025.

Tan, Mingxing et Quoc V. Le. 2020. « EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks ». In *Proceedings of the 36th International Conference on Machine Learning*. (11 septembre 2020). arXiv. <http://arxiv.org/abs/1905.11946>.

Tan, Zhiying, Yan Ji, Zhongwen Fei, Xiaobin Xu et Baolai Zhao. 2020. « Image-Based Scratch Detection by Fuzzy Clustering and Morphological Features ». *Applied Sciences*, vol. 10, nº 18, p. 6490. <https://doi.org/10.3390/app10186490>.

Tang, Xinyi, Yujia Yang, Li Huang et Li Qiu. 2022. « The Application of Texture Feature Analysis of Rectus Femoris Based on Local Binary Pattern (LBP) Combined With Gray-Level Co-Occurrence Matrix (GLCM) in Sarcopenia ». *Journal of Ultrasound in Medicine*, vol. 41, nº 9, p. 2169-2179. <https://doi.org/10.1002/jum.15896>.

Tomczak, L., V. Mosorov, D. Sankowski et J. Nowakowski. 2007. « Image Defect Detection Methods for Visual Inspection Systems ». In *2007 9th International Conference - The Experience of Designing and Applications of CAD Systems in Microelectronics*. (février 2007), p. 454-456. <https://doi.org/10.1109/CADSM.2007.4297617>.

Tomczak, Lukasz, Volodymyr Mosorov et Dominik Sankowski. 2006. « Texture Defect Detection with Non-Supervised Clustering ». In *2006 International Conference - Modern Problems of Radio Engineering, Telecommunications, and Computer Science*. (février 2006), p. 266-268. <https://doi.org/10.1109/TCSET.2006.4404516>.

UEFT. [s d]. « Ultra Electronics Forensic Technology ». <https://www.ultra-forensictechnology.com/en>. Consulté le 20 juin 2022.

Yang, Bo, Xiao Fu, Nicholas D. Sidiropoulos et Mingyi Hong. 2017. *Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering*. <http://arxiv.org/abs/1610.04794>. Consulté le 18 juillet 2023.

Yang, Guosheng, Yanli Hou et Chunyan Huang. 2004. « Texture segmentation algorithm based on wavelet transform and kd-tree clustering ». In *IEEE Conference on Robotics, Automation and Mechatronics, 2004.* (décembre 2004), p. 987-990 vol.2. <https://doi.org/10.1109/RAMECH.2004.1438053>.

Yang, Min, Dong-yun Li, Li Mou et Wei-dong Wang. 2008. « Striation Patterns Classification of Tool Marks Based on Extended Fractal Analysis ». In *2008 Chinese Conference on Pattern Recognition*. (octobre 2008), p. 1-5. <https://doi.org/10.1109/CCPR.2008.93>.

Yang, Min et Li Mou. 2009. « Striation Patterns Classification of Tool Marks Based on Morphological Structure Features ». In *2009 2nd International Congress on Image and Signal Processing*. (octobre 2009), p. 1-5. <https://doi.org/10.1109/CISP.2009.5301290>.

Yıldırım, Soner. 2020. « DBSCAN Clustering — Explained ». In *Medium - Towards Data Science*. <https://towardsdatascience.com/dbscan-clustering-explained-97556a2ad556>. Consulté le 12 décembre 2022.

Zheng, X., J. Soons, T. V. Vorburger, J. Song, T. Renegar et R. Thompson. 2014. « Applications of surface metrology in firearm identification ». *Surface Topography: Metrology and Properties*, vol. 2, nº 1, p. 014012. <https://doi.org/10.1088/2051-672X/2/1/014012>.

Zhu, Jialing, Rongjing Hong, Ashraf Uz Zaman Robin et Hao Zhang. 2022. « Deep-learning based method for breech face comparisons ». In *2022 The 6th International Conference on Machine Learning and Soft Computing (ICMLSC)*. (Haikou China, 15 janvier 2022), p. 15-19. ACM. <https://doi.org/10.1145/3523150.3523153>.