UCLouvain
FUCaM MONS

LOUVAIN
School of Management

Louvain School of Management

Obsolescence des détecteurs de deepfakes GAN-trained face aux modèles de diffusion : gradient systématique de détectabilité selon la distance architecturale et récupération pragmatique de robustesse par stratégies d'ensemble à poids gelés

Auteur : Ducarme Maxime
Promoteur(s) : Vande Kerckhove Corentin
Année académique 2025-2026
Master [120] Ingénieur de gestion, à finalité spécialisée

Déclaration d'utilisation IA générative

DECLARATION

DECLARATION D'UTILISATION D'IA GÉNÉRATIVE

Confirmez soit les quatre premières déclarations, soit, le cas échéant, la dernière :

Déclarations relatives à l'usage responsable de l'IA générative	**Réponse**
Le contenu de mon/notre TFE est mon/notre production originale, même si j'ai/nous avons utilisé l'IA générative pour m'/nous aider à le préparer.	☒ oui ☐ non
Je/nous maîtrise/maîtrisons les théories, outils et méthodes que j'utilise/nous utilisons pour le mémoire.	☒ oui ☐ non
J'ai/nous avons vérifié que le contenu présenté dans le mémoire est exact et valide, et qu'il ne contient ni hallucination, ni fausse référence, etc.	☒ oui ☐ non
J'atteste/nous attestons maîtriser le résultat final et être capable(s) de l'expliquer et de répondre aux questions à son sujet.	☒ oui ☐ non
J'ai/nous avons pris toutes les mesures nécessaires afin de ne pas donner accès à des informations ou données confidentielles à des outils d'IA générative	☒ oui ☐ non

SIGNATURE

Signez votre déclaration

	Auteur Nom de famille, prénom(s)	**Date**	**Signature**
1	Ducarme, Maxime	27/05/26	

NOTE

Pour rédiger ce document, Deepl et Claude Anthropic ont été utilisés afin de suggérer des formulations, mais le texte est original, à l'exception d'extraits explicitement mentionnés comme venant d'autres sources, mentionnées en notes de bas de page. La bibliographie a été réalisée avec Zotero.

Résumé

Les détecteurs de deepfakes entraînés sur des images issues de réseaux antagonistes génératifs (GAN) perdent significativement en efficacité face aux modèles de diffusion (DM), dont les artefacts diffèrent fondamentalement. Si cette obsolescence a été documentée, deux questions demeurent ouvertes : existe-t-il un lien entre l'architecture des modèles de diffusion et leur niveau de détectabilité ? Et peut-on améliorer les performances sans réentraînement coûteux, à l'aide de modèles d'apprentissage d'ensemble ? Cette recherche évalue quatre détecteurs GAN-trained (Meso4, XceptionNet, UCF et F3Net) sur 7 020 images du dataset DF40, combinant images réelles et deepfakes générés par quatre architectures sélectionnées pour leur distance croissante avec les GAN : ddim (convolutionnels), MidJourney (Latent Diffusion Model), DiT et SiT (Transformers purs). Trois stratégies d'ensemble sont testées via un protocole expérimental rigoureux appliquant un split vidéo-level strict.

Les résultats observés démontrent l'existence d'un gradient systématique de détectabilité directement corrélé à la distance architecturale entre les DM, bien que partiellement validé au niveau intra-famille Transformer. En effet, les architectures convolutionnelles (ddim et MidJourney) atteignent des AUC entre 0.75 et 0.86 alors que les Transformers (DiT et SiT) chutent entre 0.59 et 0.72. Concernant les détecteurs, Meso4 présente une obsolescence totale avec des prédictions équivalentes au hasard (AUC 0.502). XceptionNet, UCF et F3Net subissent, quant à eux, une dégradation graduée contrôlée (AUC 0.751-0.770). Le méta-learner (scénario C) atteint les meilleures performances (AUC 0.788), soit un gain de 1.88 pp sur XceptionNet, résultant d'un effet de recalibration exploitant la moyenne P(FAKE) élevée de Meso4 et de la complémentarité partielle entre détecteurs naïfs, spatiaux et fréquentiels

Ce gain de 1.88 point d'AUC est accessible en moins de 5 minutes (contre des dizaines d'heures pour un réentraînement complet), rendant cette solution pertinente lorsque les deepfakes proviennent majoritairement d'architectures convolutionnelles. Ces résultats offrent aux organisations un cadre de décision directement actionnable : une exposition majoritairement convolutionnelle (MidJourney, ddim) justifie le recours à une approche d'ensemble légère, suffisante et quasi-immédiate ; à l'inverse, une exposition aux architectures Transformer (DiT, SiT) ou un contexte à faible tolérance aux faux négatifs — vérification d'identité, opérations bancaires — rend le réentraînement structurellement nécessaire, le plancher incompressible de 21.7% constituant une limite que les méthodes d'ensemble seules ne peuvent franchir.

Remerciements

Avant de réaliser la lecture de ce mémoire, je tiens à remercier les différentes personnes ayant contribué, de près ou de loin, à la réalisation de celui-ci.

Je souhaiterais dans un premier temps remercier mon promoteur, Monsieur Vande Kerckhove Corentin, pour son encadrement rigoureux, sa disponibilité et son engouement autour de mon projet. Son regard critique et ses conseils avisés m'ont permis de clarifier ma méthodologie de travail et mon orientation expérimentale.

Je tiens également à exprimer ma reconnaissance envers la Louvain School of Management pour la formation académique active et qualitative que j'ai eu l'opportunité de suivre durant ces 5 dernières années.

Enfin, j'aimerais remercier mes parents, Laurence Vercaigne et Thierry Ducarme pour le soutien engagé et continu dont ils ont fait preuve tout au long de mon parcours universitaire. Leur courage et leur bienveillance m'ont permis de réaliser ce mémoire dans les meilleures conditions possibles.

Table des matières

Liste des tableaux

Liste des figures

Introduction générale

Dans un monde en constante évolution technologique, l'intelligence artificielle s'est imposée comme un élément fondamental au développement de notre société. Sa présence croissante dans les domaines d'activité a transformé notre manière de travailler, de communiquer, de créer. Parmi ses intégrations, nous retrouvons un processus de Deep Learning de plus en plus populaire : la génération de deepfakes. Selon l'article 3(60) de l'AI Act (Artificial Intelligence Act, 2024), un deepfake est un « contenu image, audio ou vidéo généré ou manipulé par l'intelligence artificielle qui ressemble à des personnes, des objets, des lieux, des entités ou des événements existants et qui pourrait paraître faussement authentique ou véridique à une personne ». La diffusion massive de ces contenus trompeurs nécessite une véritable prudence quant à notre interprétation des informations numériques (Rana et al., 2022).

Au-delà de l'impact technologique, les deepfakes représentent un danger majeur pour les sphères économiques, politiques et sociales. En 2019 par exemple, un groupe de criminels a utilisé un logiciel basé sur l'intelligence artificielle pour imiter la voix d'un PDG et exiger un virement frauduleux de 220 000 € à une entreprise britannique (Stupp, 2019). Un exemple plus récent fût à Hong Kong, en 2024, où une multinationale a perdu 26 millions de dollars suite à une visioconférence où tous les participants, sauf un, étaient des deepfakes (Inter, 2024). Ces cas illustrent la puissance de destruction économique et sociales des deepfakes. Une entreprise victime perd non seulement des ressources financières, mais voit également sa réputation et sa crédibilité se détériorer. Sur la sphère politique, il est devenu assez courant de visionner des contenus trompeurs visant des politiciens dans le but d'influencer les convictions du public numérique (Sharma et al., 2024).

Ce danger est d'autant plus présent que l'accessibilité à ces outils de génération croît à une vitesse impressionnante. Depuis 2022, la démocratisation de ces outils au grand public s'est accélérée par la mise à disposition de Stable Diffusion en open source, le lancement de MidJourney et DALL-E, et l'adoption massive des IA génératives textuelles comme ChatGPT pour l'exemple (Rombach et al., 2022). Ainsi, là où la génération de deepfakes convaincants nécessitait auparavant des connaissances et compétences en apprentissage automatique, elle ne demande désormais plus qu'une interface web et quelques secondes de traitement, faisant d'elle une solution de premier recours bien ancrée dans les méthodes de travail contemporaines.

Afin de répondre à cette menace, la détection automatique de deepfakes est devenue un enjeu fondamental de la lutte contre la désinformation et la fraude numérique. Historiquement, les chercheurs ont développé des détecteurs spécialisés dans la détection des deepfakes produits par les Réseaux Antagonistes Génératifs (GAN). Introduit par Goodfellow en 2014, ces générateurs fonctionnent sur un principe d'apprentissage contradictoire entre deux réseaux de neurones, laissant des traces caractéristiques que les détecteurs forensiques ont appris à exploiter. Durant plusieurs années, ces détecteurs ont montré des performances impressionnantes, atteignant des taux de détection parfois supérieurs à 95% sur des jeux de données standards comme FaceForensics++ (Rossler et al., 2019). Cependant, le paysage technologique a connu un bouleversement majeur avec l'émergence des modèles de diffusion, popularisés par Ho et al. en 2020. Ces modèles fonctionnent sur un principe de débruitage itératif progressif, produisant des deepfakes d'un réalisme nettement supérieur aux GAN. En effet, les recherches de Liu et al. (2025) démontrent que contrairement aux GAN qui compressent l'information, les DM travaillent souvent sur un espace latent de taille réelle, préservant ainsi une finesse de détails supérieure. Cette même étude conclut que les méthodes de détection GAN deviennent largement inefficaces face aux nouvelles méthodes de générations. Ce constat révèle un véritable bouleversement au sein de la sphère scientifique. En effet, élaborer et entraîner de tout nouveaux détecteurs engendrent des ressources financières, temporelles, computationnelles, humaines et environnementales non-négligeables (Doloriel et al., 2026; Fontana et al., 2025). De plus, selon une étude de Ricker et al. (2024), les détecteurs conçus sur GAN ne généralisent pas bien sur les DM sans réentrainement et l'ère de la détection de ces derniers n'en est qu'à ses débuts.

Face à ce constat, une solution évidente consisterait à entraîner de nouveaux détecteurs directement sur des images issues de modèles de diffusion. Cependant, cette approche se heurte à des contraintes opérationnelles considérables : le réentraînement complet d'un détecteur de deepfakes nécessite entre 52 et 74 heures de calcul GPU et représente un coût estimé entre 500 et 2 000€ selon l'infrastructure utilisée (Fontana et al., 2025; Ricker et al., 2024). Dans un contexte où de nouvelles architectures de diffusion émergent à un rythme mensuel, ce cycle de réentraînement devient structurellement incompatible avec la vitesse d'évolution de la menace. Une alternative pragmatique consiste à combiner des détecteurs GAN-trained existants au sein d'approches d'ensemble à poids gelés — c'est-à-dire des stratégies qui agrègent les prédictions de plusieurs détecteurs sans modifier leurs paramètres internes ni nécessiter de réentraînement, exploitant uniquement la complémentarité de leurs signaux respectifs. Cette contrainte de poids

gelés simule la réalité opérationnelle des organisations qui ne peuvent se permettre un réinvestissement computationnel face à chaque nouvelle menace.

Malgré la découverte de cette problématique d'obsolescence, deux questions fondamentales restent ouvertes et constituent le cœur de ce mémoire. Premièrement, toutes les architectures DM posent-elles les mêmes difficultés aux détecteurs GAN ou bien existe-t-il un gradient systématique de détectabilité directement corrélé à la distance architecturale qu'elles possèdent par rapport au(x) GAN utilisé(s) pour l'entraînement ? Deuxièmement, dans quelle mesure l'élaboration d'approches d'ensemble à poids gelés peut-elle améliorer les performances de détection et leur robustesse face à de nouvelles architectures ? Les réponses à ces questions permettront à la recherche d'être en mesure de déterminer, selon le contexte architectural, si la détection de deepfakes issus de modèles de diffusion nécessite le déploiement et/ou l'(le) (ré)entrainement coûteux de nouveaux détecteurs spécialisés, ou si la combinaison de détecteurs GAN constitue toujours une option fiable et efficace. Notre question de recherche centrale s'articule de la façon suivante : **comment différentes architectures de modèles de diffusion impactent-elles la performance des détecteurs GAN-trained, et une approche d'ensemble à poids gelés peut-elle pragmatiquement récupérer de la robustesse sans réentraînement** ?

Afin de structurer cette étude, le présent mémoire s'organise en cinq chapitres complémentaires. Le chapitre 1 concerne la revue de littérature et pose les fondements théoriques nécessaires à la compréhension. Celui-ci présente l'évolution des modèles génératifs, allant des GAN aux modèles de diffusion, les principaux détecteurs développés par la communauté scientifique, ainsi qu'une synthèse condensée concernant l'état actuel de la recherche. Le chapitre 2 expose la méthodologie employée, détaillant et justifiant la sélection des quatre détecteurs testés, la constitution du jeu de données utilisé, l'explication des scénarios d'ensemble élaborés, ainsi que les protocoles expérimentaux utilisés rigoureusement afin de garantir la validité des résultats. Le chapitre 3 présente les résultats empiriques issus de nos expériences, quantifiant précisément le gradient de détectabilité entre méthodes de diffusion ainsi que la performance des trois scénarios d'ensemble testés. Il se conclut par une synthèse récapitulative de l'évaluation expérimentale. Le chapitre 4 interprète les mécanismes sous-jacents aux performances et positionne nos résultats par rapport à la recherche scientifique tout en exposant les limitations, les perspectives d'améliorations, et les implications managériales liées à cette recherche. Enfin, le chapitre 5 conclut en synthétisant les contributions scientifiques majeures de cette recherche et en apportant une réponse finale à notre question centrale.

Chapitre 1 : Revue de littérature

1.1 Evolution des modèles génératifs

1.1.1 Réseaux génératifs antagonistes (GAN)

En 2012, après une première période dominée par l'apprentissage automatique supervisé où les algorithmes apprenaient sur des exemples étiquetés, le deep learning s'est imposé comme une révolution marquante permettant aux réseaux de neurones profonds d'atteindre des performances inégalées sur des tâches complexes de reconnaissance visuelle et de traitement du langage. Cette évolution est en grande partie due à la disponibilité croissante de données massives et à la démocratisation des outils de traitement graphique (GPU) permettant leur traitement. Les réseaux génératifs antagonistes s'inscrivent dans cette phase d'évolution (LeCun et al., 2015).

En effet, introduits par Goodfellow et al. en 2014, les GAN ont représenté la technologie principale de création de contenus synthétiques hyperréalistes pendant près d'une décennie. Le fonctionnement des GAN, présenté par la Figure 1, repose sur une compétition itérative entre deux réseaux de neurones : un générateur et un discriminateur (Goodfellow et al., 2014). À chaque cycle, le générateur transforme un vecteur de bruit aléatoire en image synthétique dans le but de tromper le discriminateur, celui-ci ayant pour rôle de discerner le vrai du faux au sein d'un ensemble de données. Ce processus adversarial peut être assimilé à un problème d'optimisation minimax où le générateur tente de maximiser la probabilité que le discriminateur classifie incorrectement ses productions comme réelles (Liu et al., 2025).

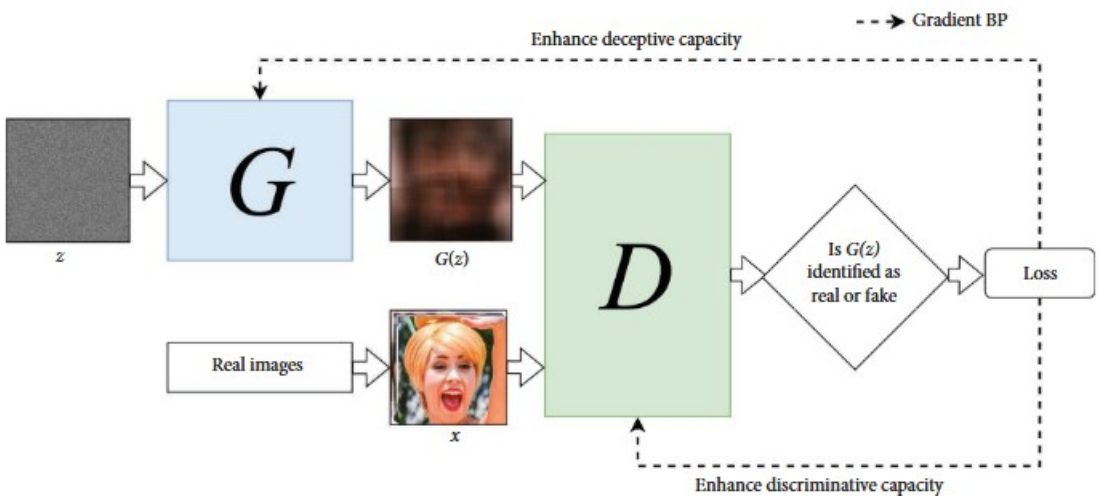

Figure 1 : Architecture de base d'un GAN (Liu et al., 2025)

Cependant, les images issues de GAN laissent des artefacts que les détecteurs ont appris à identifier. Lors de la génération, le modèle générateur effectue une projection d'un code latent compressé vers une sortie de dimension supérieure telle qu'une image. Ce code est délibérément limité à une dimension inférieure à celle de sortie. Cette compression oblige le générateur à apprendre une représentation compressée de la distribution des données d'apprentissage et se traduit en perte d'informations identifiables sur le spectre fréquentiel (Liu et al., 2025; Ricker et al., 2024).

1.1.2 Modèles de diffusion (DM)

Les modèles de diffusion représentent un changement fondamental dans la façon de générer des deepfakes. En effet, du fait de leur processus de débruitage itératif présenté en Figure 2, les DM produisent des résultats de meilleure qualité que les GAN tout en laissant des artefacts différents et moins facilement détectables. Le débruitage itératif se déroule en deux étapes : d'abord, on ajoute du bruit gaussien à une image réelle de manière progressive jusqu'à obtenir du bruit pur, puis on entraîne un réseau de neurones à inverser ce processus pour générer une nouvelle image en partant de ce bruit (Liu et al., 2025).

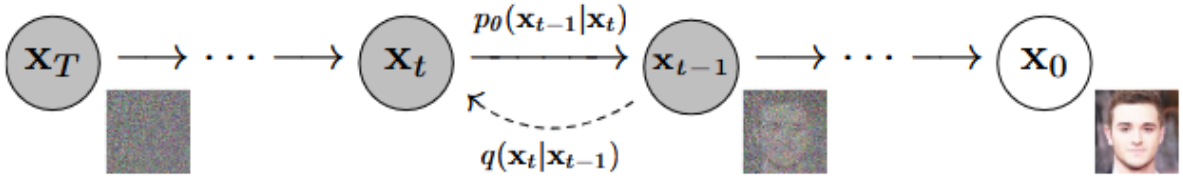

Figure 2 : Processus forward et reverse d'un modèle de diffusion (Ho et al., 2020)

A l'inverse des GAN, les DM sont beaucoup plus volumineux en termes de paramètres et de taille de modèles. Certains modèles de diffusion, notamment les architectures pixel-space comme DDPM, travaillent directement à l'échelle de l'image finale sans imposer de compression latente drastique. D'autres, comme les Latent Diffusion Models (MidJourney), utilisent une compression via auto-encodeur variationnel (VAE), mais celle-ci reste moins contraignante que la compression imposée par les GAN. Cette flexibilité architecturale permet de mieux préserver les détails fins au cours du processus et de fournir des résultats de meilleure qualité (Liu et al., 2025). Cependant, Ricker et al. (2024) observent dans leur étude une sous-estimation systématique de la densité spectrale, qu'ils attribuent directement à la fonction de perte des DM.

1.1.3 Transformer

Parmi les DM figurent une architecture structurellement différente des réseaux de neurones convolutionnels (CNN, Convolutional Neural Network) et souvent plus efficace : l'architecture Transformer. Développée en 2017, celle-ci repose exclusivement sur un mécanisme d'attention, se passant totalement des principes de récurrence ou de convolutions (Vaswani et al., 2017).

Le mécanisme d'attention fonctionne sur un système de mise en relation où le modèle calcule le degré de pertinence de chaque élément d'une séquence avec tous les autres éléments simultanément. Plus concrètement, chaque paire d'éléments se voit attribuer un poids d'importance permettant au modèle de déterminer quelles parties de l'entrée doivent influencer la représentation d'une région donnée. Ainsi, contrairement aux CNN qui analysent une image par petites zones locales, l'attention permet de connecter directement des zones éloignées de l'image à partir des premières couches du réseau (Vaswani et al., 2017).

L'émergence de cette architecture au sein de la génération d'images s'est concrétisée avec les Vision Transformers (Dosovitskiy et al., 2021). Ceux-ci découpent l'image en patches réguliers qui sont traités comme une séquence de tokens. Le Diffusion Transformer (DiT) (Peebles & Xie, 2023) intègre ce principe dans un processus de diffusion en deux étapes : un VAE compresse d'abord l'image dans un espace latent, puis ces représentations sont divisées en patches sur lesquels le Transformer applique son mécanisme d'attention pour apprendre à débruiter progressivement l'image.

Cette différence architecturale avec les CNN soulève une question centrale concernant la capacité des détecteurs à généraliser entre ces deux familles. A l'heure actuelle, peu de recherches permettent de quantifier l'ampleur d'une potentielle perte de performance d'un détecteur entraîné sur l'une de ces familles lorsqu'il est testé sur l'autre.

1.2 Obsolescence des GAN-trained sur les modèles de diffusion

Les travaux scientifiques récents ont prouvé que les détecteurs entraînés sur des images produites par des GAN (GAN-trained) perdent significativement en performance lorsqu'ils sont confrontés à des images issues de DM (Choi & Woo, 2025; Corvi et al., 2022). Cette détection d'obsolescence s'explique principalement par la nature fondamentalement différente des artefacts laissés par chacun des processus génératifs. Cette observation représente une

découverte majeure dans le domaine de la détection et soulève la problématique relative à la capacité de généralisation des détecteurs.

Les recherches scientifiques liées à cette problématique sont assez controversées. D'un côté, les travaux de Ricker et al. (2024) démontrent que les détecteurs entraînés sur des images de diffusion détectent en moyenne 94.5% des images issues de GAN alors que les détecteurs entraînés sur GAN ne détectent que 26.4% des images DM. Ces résultats suggèrent que les détecteurs DM généralisent remarquablement bien contrairement aux détecteurs GAN, sans réentraînement quelconque. Selon eux, les artefacts laissés par les GAN sont tellement basiques qu'ils sont facilement reconnus par les premières couches de convolutions des détecteurs réseaux de neurones (Ricker et al., 2024).

D'un autre côté, l'étude réalisée en 2025 par Choi et Woo vient nuancer ces asymétries de généralisation entre GAN et DM. Celle-ci suggère que les conclusions sur la généralisation des DM vers GAN pourraient fortement dépendre des choix de datasets, détecteurs et protocoles expérimentaux. En effet, leur analyse démontre que les modèles EfficientNet et XceptionNet entraînés sur des données générées par des GAN surpassent systématiquement ces mêmes modèles entraînés sur des DM en matière de généralisation inter-famille. Cette supériorité indique que les données GAN fournissent des caractéristiques globales permettant d'améliorer les capacités de généralisation inter-familles des détecteurs sur ces deux modèles. L'étude insiste toutefois sur le fait que les performances peuvent fortement varier selon les datasets et détecteurs choisis (Choi & Woo, 2025).

Concernant la capacité de réentraînement, Ricker et ses collaborateurs ont montré que le réentraînement des détecteurs GAN sur images DM permet d'obtenir des performances de détection significativement précises, et inversement. Cela nécessite toutefois des ressources considérables (Ricker et al., 2024).

De manière générale, les performances restent assez logiquement bien supérieures lorsque l'on essaie de détecter des images issues de la même famille que celle d'entraînement des détecteurs. Sur ce constat, Guarnera et ses collaborateurs ont développé en 2024 une approche multi-niveau permettant de distinguer les images générées par GAN de celles produites par DM. Cette approche permettrait d'appliquer directement le détecteur entraîné sur la bonne famille. Celle-ci démontre qu'il est techniquement possible de classifier ces deux familles avec plus de 97% de précision. Cependant, cette performance nécessite le développement d'architecture

spécifiquement développée pour cette tâche, les détecteurs GAN-trained et DM-trained actuels n'étant pas capables de faire la différence (Guarnera et al., 2024).

1.3 Limites identifiées dans la littérature

Malgré les avancées récentes dans la caractérisation de l'obsolescence des détecteurs GAN lorsqu'ils sont appliqués en dehors de leur base d'entraînement, deux questions cruciales restent largement ouvertes dans la littérature scientifique. Premièrement, la majorité des travaux existants traitent les modèles de diffusion comme une catégorie homogène, sans distinguer entre les différentes architectures qui composent cette famille. Or, les modèles de diffusion couvrent un spectre architectural large, des méthodes à débruitage convolutionnel comme ddim jusqu'aux architectures Transformers pures comme DiT et SiT. La question de savoir si toutes ces architectures posent le même niveau de difficulté aux détecteurs, ou s'il existe un gradient systématique de détectabilité corrélé à leur distance architecturale par rapport aux GAN, n'a pas été explorée de manière systématique.

Deuxièmement, bien que Ricker et al. aient démontré que le réentraînement des détecteurs sur des données de diffusion permet une détection presque parfaite (Ricker et al., 2024), cette solution nécessite des ressources considérables d'un point de vue temporel et computationnel mais également une forte complexité pour concevoir les ensembles de données adéquats. En effet, des études récentes démontrent que le réentraînement complet d'un détecteur de deepfakes de taille moyenne nécessite entre 52 et 74 heures de calcul GPU pour un seul cycle d'adaptation (Fontana et al., 2025). Le coût associé est estimé entre 500 et 2000€ selon l'architecture et l'infrastructure utilisée (Ricker et al., 2024). A titre comparatif, les approches d'apprentissage continu permettent d'atteindre des performances comparables en environ 5 à 18 minutes (Fontana et al., 2025). Bien entendu, ces données peuvent varier selon le contexte d'expérimentation. L'ordre de grandeur de ces tendances atteste toutefois de l'efficacité computationnelle des approches d'ensemble.

Cette barrière économique et temporelle limite considérablement la réactivité opérationnelle des organisations face à l'émergence de nouvelles techniques de génération. Dans un contexte où les modèles de diffusion évoluent à un rythme mensuel, avec des architectures comme DiT et SiT introduisant des ruptures architecturales majeures (Peebles & Xie, 2023), le cycle de réentraînement complet devient structurellement incompatible avec la vitesse d'évolution de la menace. De plus, la perspective "Green AI" (Doloriel et al., 2026) souligne l'impact

environnemental non négligeable de ces réentraînements fréquents, renforçant la nécessité de solutions d'adaptation légères et efficaces. Ainsi, dans un contexte opérationnel où les organisations impactées par l'émergence des deepfakes doivent s'adapter rapidement aux nouvelles menaces, la question de savoir comment améliorer les performances sans devoir passer par une phase de réentraînement demeure largement ouverte.

1.4 Approches d'ensemble : état de l'art

Afin de faire face aux problématiques de ressources liées aux réentraînements des détecteurs de deepfakes, l'élaboration des modèles d'ensemble est apparue comme une solution logique et potentiellement efficace au sein de la recherche scientifique. L'intuition sous-jacente est relativement simple : si différents détecteurs captent des artefacts complémentaires (spatiaux, temporels, fréquentiels, etc.), leur combinaison devrait théoriquement compenser les faiblesses de chacun et fournir des résultats avec des performances supérieures. Les résultats empiriques de la littérature tendent à confirmer cette hypothèse. Wahab et ses collaborateurs ont observé en 2025 que les prédictions d'ensemble offraient une stabilité bien supérieure aux modèles individuels, dont les performances variaient considérablement selon les datasets testés (Wahab et al., 2025).

Les stratégies d'agrégation des prédictions se divisent en deux catégories principales de modèles d'ensemble. La première concerne les agrégations fixes et ne nécessite aucun entraînement. Celles-ci combinent les sorties des détecteurs via des opérations mathématiques prédéfinies, par exemple une moyenne arithmétique simple ou une moyenne pondérée. La deuxième concerne les combinaisons apprises. Souvent des régressions logistiques, celles-ci entraînent le modèle sur base des scores des détecteurs de base afin d'obtenir les coefficients de pondération optimaux (Staněk et al., 2026). Cette deuxième approche est souvent désignée sous le nom de méta-learning et a démontré son efficacité au sein de différents contextes de détection de deepfakes. Par exemple, Naskar et ses collaborateurs ont appliqué un méta-learner entraîné sur les prédictions de modèles de base pour améliorer la généralisation cross-dataset entre CelebDF et FaceForensics++ (Naskar et al., 2024).

Cependant, la plupart de ces travaux sur le méta-learning identifiés se concentrent sur la généralisation entre datasets au sein d'une même famille de modèles génératifs, par exemple d'un ensemble GAN vers un autre ensemble GAN. Le contexte inter-familial pose quant à lui un défi plus relevé. En effet, il ne suffit pas de généraliser vers de nouvelles données, mais vers

une technologie différente dont les artefacts sont structurellement distincts. De plus, la question de savoir si un méta-learner entraîné uniquement sur les scores de détecteurs GAN-trained à poids gelés peut récupérer de la robustesse face aux modèles de diffusion reste, à notre connaissance, inexplorée au sein de la littérature. L'enjeu de cette problématique est crucial d'un point de vue opérationnel : un méta-learner entraîné sur scores de confiance ne manipule jamais les images brutes, ne nécessite que quelques centaines d'exemples pour calibration, et représente un coût computationnel négligeable comparé au réentraînement des détecteurs de base.

1.5 Positionnement de cette recherche

Le positionnement de cette recherche a pour objectif de combler deux lacunes principales identifiées dans la littérature scientifique.

Premièrement, si la perte de performance significative des détecteurs GAN-trained sur les images DM a déjà pu être validée (Choi & Woo, 2025; Guarnera et al., 2024; Ricker et al., 2024), une analyse plus spécifique orientée sur le type d'architecture de ces images reste une question ouverte. En effet, les recherches portées sur la généralisation des GAN aux DM ont souvent considéré les modèles de diffusion comme une structure uniforme, ce traitement étant justifié par le fait que cette famille est relativement récente. Cependant, la littérature affirme que la famille des modèles de diffusion comprend un panel de types d'architecture multiples et diversifiés, allant des méthodes à débruitage convolutionnel comme ddim aux générateurs commerciaux optimisés comme MidJourney, jusqu'aux architectures Transformers pures comme DiT et SiT qui s'éloignent radicalement des biais convolutifs des GAN. Cette recherche permet d'évaluer l'impact des différentes structures relatives aux modèles de diffusion sur la capacité de généralisation des détecteurs GAN-trained. Ainsi, nous pouvons formuler l'hypothèse que, corrélé à la distance architecturale avec les mécanismes GAN, un gradient systématique de détectabilité pourrait émerger.

Deuxièmement, cette recherche évalue la capacité des stratégies d'ensemble à poids gelés à améliorer la généralisation des détecteurs GAN-trained sur les DM. Les modèles génératifs évoluent à grande vitesse et, l'entraînement et le développement de détecteurs opérationnels étant extrêmement coûteux en ressources computationnelles, temporelles et financières, les organisations ciblées par la menace des deepfakes doivent réagir extrêmement rapidement. Afin de répondre à ce besoin, le développement de techniques d'apprentissage d'ensemble a déjà fait

ses preuves sur des contextes intra-familiaux (Rana et al., 2022; Staněk et al., 2026). Cette recherche permet d'évaluer l'impact de ces stratégies sur un contexte inter-familial GAN vers DM. Trois stratégies sont comparées : moyenne arithmétique simple, moyenne pondérée par précision (accuracy), et méta-learner logistique entraîné sur les scores de confiance des détecteurs gelés. L'objectif n'est pas de démontrer qu'une telle approche surpasse un réentraînement complet — ce serait irréaliste — mais de quantifier précisément le gain de robustesse réellement accessible sans réentraînement, permettant aux organisations d'évaluer si des solutions légères suffisent ou si un réinvestissement dans le réentraînement est nécessaire.

Conformément aux standards de la recherche en détection de deepfakes, notre étude se concentre exclusivement sur des images de visages humains. Cette convention scientifique s'explique par le fait que les visages humains synthétiques représentent la catégorie la plus critique d'un point de vue opérationnel : fraude d'identité, désinformation politique, harcèlement digital. La disponibilité des datasets et la facilité de preprocessing (extraction et alignement facial) justifie également cette prise de position. Toutefois, cette limitation implique que nos conclusions ne sont pas généralisables aux deepfakes issus d'autres catégories visuelles.

Cette recherche voit également apparaître les notions d'obsolescence et de dégradation graduée. Le terme obsolescence est attribué lorsque les performances sont statistiquement indistinguables du hasard, soit une AUC ≈ 0.5. Le terme dégradation graduée concerne les détecteurs qui conservent une capacité discriminante supérieure au hasard malgré une perte en performance par rapport à son AUC en intra-domaine.

Chapitre 2 : Méthodologie

Ce chapitre présente la démarche méthodologique employée afin d'effectuer les expérimentations relatives à la question de recherche de cette étude. Celui-ci sera construit autour de 6 axes élémentaires : la sélection de l'environnement de travail, la constitution du jeu de données, la sélection des détecteurs GAN-trained, l'élaboration des stratégies d'ensemble, l'explication des métriques utilisées et enfin la présentation du flux expérimental suivi.

Cette approche permet de répondre à notre question de recherche tout en simulant le contexte le plus opérationnel possible à notre échelle, afin que les résultats obtenus soient réellement comparables avec ceux des organisations sur le terrain.

2.1 Framework d'évaluation : DeepfakeBench

Afin de mener ces expérimentations, notre recherche s'appuie sur un benchmark GitHub de référence nommé 'DeepfakeBench' (SCLBD, 2026) (Annexe A : .2). Ce benchmark d'évaluation standardisé est largement reconnu au sein de la sphère de détection de deepfakes. Il fournit 36 architectures de détection avec leur poids pré-entraînés et leur fichier de configuration YAML. Ces architectures ont été entraînés sur FaceForensics++ c23 contenant des images de visage réelles et synthétiques générées par GAN. Ce jeu de données figure parmi les plus fiables et complets au sein de la recherche scientifique.

Toutefois, dans le cadre de cette étude, le framework est utilisé en dehors de son pipeline natif d'entraînement et d'évaluation. Seules les parties architecturales des détecteurs ont été récupérées. Le développement du code lié à nos expérimentations a été implémenté en Python 3.10 via Google Colab avec stockage sur Google Drive (Annexe A :). Les quatre détecteurs ont été utilisés avec leurs poids FaceForensics++ c23 originaux, sans fine-tuning ni réentraînement. Le code source complet de cette étude est accessible publiquement sur GitHub[1] (Annexe A.1).

2.2 Constitution du jeu de données

2.2.1 Sélection du dataset DF40

Le jeu de données utilisé dans cette recherche provient de DF40 (DeepfakeBench v2), un dataset académique publié lors de NeurIPS 2024 et rassemblant des images générées par 40

[1] https://github.com/ducamax07/deepfake-detector-obsolescence.git

techniques de synthèses différentes. Ce dataset a été spécifiquement conçu pour évaluer la robustesse des détecteurs de deepfakes face à la diversité croissante des méthodes de génération. Le choix de ce jeu de données repose sur différents critères de sélection établis.

Premièrement, le preprocessing sur DF40 a déjà été réalisé, cela nous faisant économiser temps et énergie. Les images fournies sont des visages alignés et normalisés dans un format standard 256*256 pixels en mode RGB, éliminant ainsi tout biais lié à la résolution, la compression ou l'alignement de l'image. Cette homogénéité est cruciale afin d'éviter tout artefact parasite lié à la génération ou au preprocessing. Deuxièmement, DF40 fournit des images issues de 40 méthodes de génération de deepfakes différentes (GAN et DM). Cela permet de constituer le set de données le plus complet possible relativement à nos intentions. Dans notre cas, 4 méthodes DM sont finalement scrupuleusement sélectionnées (section 2.2.3). Troisièmement, ce dataset provient d'une source reconnue par la recherche, garantissant transparence, traçabilité de sa constitution et comparabilité fiable avec d'autres travaux.

2.2.2 Configuration définitive du dataset

La configuration finale du dataset respecte un équilibre parfait entre le nombre d'images réelles et synthétiques avec un total de 7020 images.

Les images REELLES proviennent de deux sources complémentaires conseillées par le benchmark de référence : FaceForensic++ (1858 images) et CelebDF-v2 (1652 images) (SCLBD, 2026). Cette approche bi-sources permet de diversifier le plus possible les contextes d'acquisition (éclairage, angle, colorimétrie, etc.) tout en maintenant un niveau technique élevé puisque ces deux sources sont des références validées par la recherche. Il est important de noter que ces jeux de données contiennent des séquences d'image issues de vidéos. Afin de limiter le nombre d'images ayant des caractéristiques similaires, seules 2 images issues d'une même vidéo ont pu être sélectionnées au maximum. Un échantillon aléatoire d'images réelles est disponible sous l'annexe B.3 : D5 Inspection visuelle.

Les images FAKE proviennent de 4 modèles de diffusion sélectionnés rigoureusement parmi les 40 méthodes de génération du dataset DF40. La première étape du processus de sélection consiste en l'élimination de toutes les méthodes de génération ne faisant pas partie de la famille des DM. La seconde étape consiste en un tri visuel rapide afin d'éliminer les modèles générant des images de très mauvaises qualités ou comportant des artefacts irréalistes évidents. La troisième étape consiste en une sélection définitive sur base de deux critères : la diversité

architecturale des modèles allant des processus convolutionnels proches des GAN (ddim) jusqu'aux architectures Transformers pures (SiT, DiT), et la pertinence opérationnelle afin d'avoir des modèles réellement répandus dans les contextes réels (MidJourney, ddim) et académiques (SiT, DiT). Ainsi, les quatre modèles de diffusion suivants ont été choisi : ddim, MidJourney, SiT et DiT. Pour des raisons de disponibilité des données dans DF40, les modèles à architecture convolutionnelle (1,300 images ddim et 1,594 images MidJourney) sont davantage représentés que les architectures Transformers pures (DiT : 358 images ; SiT : 258 images). Ce déséquilibre n'est pas considéré comme un biais méthodologique au sein de notre étude, mais plutôt comme un reflet de la menace opérationnelle réelle en 2026. MidJourney et ddim, par leur accessibilité (interface web, coût raisonnable, pas d'expertise ML requise), constituent les méthodes de génération de deepfakes les plus répandues parmi les acteurs malveillants. À l'inverse, DiT et SiT demeurent principalement confinés à la recherche académique en raison de leur complexité technique. Un échantillon aléatoire d'images synthétiques est disponible sous l'annexe B.3 : D5 Inspection visuelle.

La Figure 3 illustre l'équilibre entre les classes FAKE et REAL, et décompose leur répartition.

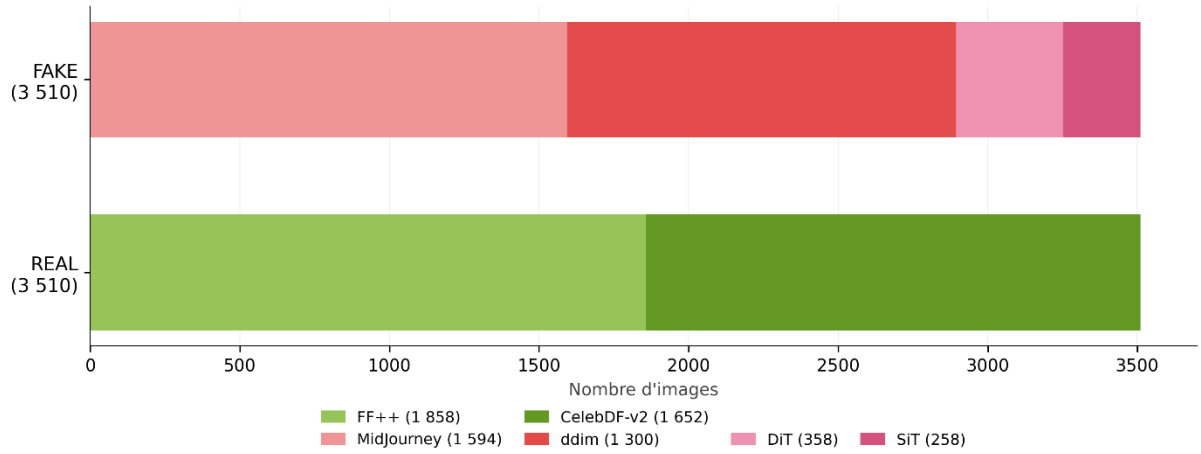

Figure 3 : Composition du dataset par classe et source de données

2.2.3 Présentation des quatre architectures de diffusion sélectionnées

Les quatre méthodes de diffusion retenues sont construites sur base d'architectures différentes nous permettant de tester empiriquement l'existence d'un gradient systématique de détectabilité.

Le modèle ddim (Denoising Diffusion Implicit Models) représente le modèle le plus « classique » de notre échantillon et par conséquent celui le plus proche des mécanismes GAN.

Introduit par Song et al. en 2022, ddim est une méthode d'échantillonnage accélérée qui reformule le processus génératif des modèles de diffusion en processus non-Markovien déterministe permettant de générer des images de qualité comparables aux générateurs de base DDPM en utilisant seulement 10 à 50 étapes au lieu de 1000. La différence entre ces modèles est que DDIM utilise un échantillonnage déterministe contrôlé par un paramètre η (η=0 pour déterministe pur, η=1 pour retrouver DDPM), ce qui accélère la génération de 10 à 50 fois. (Song et al., 2022). Par conséquent, il laisse des artefacts différents des modèles standards.

Le modèle MidJourney est le modèle le plus populaire de notre échantillon de par ses productions hyperréalistes et sa facilité d'accès par le grand public. MidJourney est un modèle propriétaire et peu d'informations sur son architecture spécifique sont officiellement publiées. Cependant, à l'instar de quelques processus d'optimisation post-diffusion probables, MidJourney agit selon le principe des Latent Diffusion Model (LDM). Au sein d'un processus LDM, un encodeur VAE compresse dans un premier temps l'image dans un espace réduit. Un réseau U-Net effectue ensuite un processus de débruitage itératif sur 20 à 100 étapes afin de transformer le bruit en représentation latente. Enfin, un décodeur VAE intervient pour reconstruire l'image dans une qualité remarquable (Rombach et al., 2022).

Le modèle Diffusion Transformer (DiT) est un des premiers réseaux Transformer utilisé pour la génération d'image synthétique et marque, par conséquent, une rupture architecturale majeure par rapport aux traditionnelles architecture U-Net convolutionnelles. Techniquement, DiT encode via VAE l'image avant de la projeter dans un espace latent compressé. Ensuite, cette représentation latente sera découpée en patches et traités comme une séquence de tokens. Le mécanisme d'attention, spécifique aux Transformers, permet alors de modéliser les dépendances à longue distance entre ces patches et de travailler sur l'image globale, contrairement aux convolutions qui agissent localement (Peebles & Xie, 2023). Les artefacts laissés par ce modèle diffèrent totalement de ceux laissés par les méthodes plus traditionnelles.

Le modèle Scalable Interpolant Transformer (SiT) pousse le processus Transformer encore plus loin en introduisant des mécanismes d'attention sparse et des stratégies de scaling optimisés : apprentissage en temps continu versus discret, prédiction de vélocité versus bruit, et échantillonnage déterministe versus stochastique. Cela permet notamment de traiter des images plus hautes résolutions sans faire exploser la complexité computationnelle. SiT est l'architecture la plus éloignée des GAN au sein de notre sélection. En effet, elle abandonne non seulement les convolutions des modèles de diffusion classique, mais elle optimise le mécanisme

d'attention relatif aux Transformers afin de minimiser les biais locaux et privilégier une modélisation globale (Ma et al., 2024).

Ainsi, ces quatre modèles génératifs apportent une diversité mesurée et pertinente afin de tester notre hypothèse sur le gradient de détectabilité. Afin de rendre cette notion de distance architecturale opérationnelle et reproductible, quatre critères binaires ordonnés sont retenus : la présence de convolutions locales comme mécanisme principal de traitement spatial, l'utilisation d'un espace latent compressé via VAE, l'intégration d'un mécanisme d'attention globale sur patches tokenisés, et l'utilisation d'une attention sparse. Les GAN reposant exclusivement sur des convolutions locales, chaque critère additionnel marque un éloignement croissant vis-à-vis de leurs mécanismes de génération, comme synthétisé dans le Tableau 1 ci-dessous.

Modèle	Convolutions locales	VAE	Attention globale	Attention sparse	Distance GAN
ddim	✅	❌	❌	❌	1
MidJourney	✅	✅	❌	❌	2
DiT	❌	✅	✅	❌	3
SiT	❌	✅	✅	✅	4

Tableau 1 : Opérationnalisation de la distance architecturale vis-à-vis des GAN

2.2.4 Protocole de split vidéo-level

Le dataset évoqué a été divisé en trois ensembles distincts selon le protocole de split vidéo-level afin de garantir l'absence de fuite d'information. Ainsi, nos images seront réparties entre ces trois sets de la façon suivante : 20% sur un Training Set, 40% sur un Test Set, et 40% sur un Validation Set (VS). Afin d'éviter tout biais provoqué par nos images réelles, celles issues de la même vidéo seront classées au sein du même ensemble.

L'implémentation technique du split a été réalisée de façon extrêmement rigoureuse sur base d'un tirage aléatoire avec seed=42 pour garantir la reproductibilité. La répartition finale est détaillée au sein du Tableau 2.

Ensemble	Total	%	REAL	FAKE	MidJourney	ddim	DiT	SiT
Train	1 402	20.0%	702	700	318	260	71	51
Test	2 807	40.0%	1 404	1 403	637	520	143	103
Validation	2 811	40.0%	1 404	1 407	639	520	144	104
TOTAL	7 020	100%	3 510	3 510	1 594	1 300	358	258

Tableau 2 : Répartition des 7 020 images entre Train, Test et Validation

Le Training Set est utile uniquement afin de calibrer les poids nécessaires au méta-learner du scénario d'ensemble C, et à assurer les coefficients de pondération du Scénario B. Étant donné que ces opérations ne nécessitent pas de réentraînement complet des détecteurs de base mais uniquement l'ajustement de paramètres d'agrégation légers, la taille de cet ensemble est volontairement réduite comparé aux autres.

Le Test Set est quant à lui volontairement étendu afin de maximiser la puissance statistique des évaluations (réduction de la variance des métriques). Il sera utilisé pour tous les tests et expérimentations de base tout au long du processus de développement.

Enfin, le Validation Set est strictement isolé tout au long du protocole expérimental jusqu'à la phase d'évaluation finale. Cela ayant pour but d'éviter toute forme quelconque de surajustement des paramètres et de garantir que les performances mesurées reflètent réellement la capacité de généralisation du modèle. Comme pour le Test Set, il représente 40% du dataset total.

2.2.5 Phase d'audit des données

Afin d'assurer la qualité et la pertinence des données à disposition, une phase d'audit en plusieurs dimensions est réalisée. Lorsqu'une image ne rentre pas dans les limitations définies au sein de ces dimensions, elle est retirée et remplacée par une autre image si cela est possible.

Cette phase d'audit repose sur 6 dimensions. La dimension D1 s'assure que chaque fichier image est techniquement exploitable, c'est-à-dire non-corrompu, de résolution 256*256 pixels et en colorimétrie RGB. La dimension D2 s'assure que le ratio REAL/FAKE est égal à 1:1, de sortes à avoir un équilibre parfait entre les deux classes (annexe B.1). La dimension D3 teste si les images FAKE et REAL présentent des différences significatives sur des métriques de bas niveau comme la luminosité, le contraste ou la saturation afin de justifier l'utilisation de détecteurs de représentations profondes (annexe B.2). La dimension D4 recherche d'éventuels doublons au sein du jeu de données. La dimension D5 consiste en une validation humaine manuelle sur un échantillon aléatoire de deepfakes afin de détecter d'éventuelles anomalies visuellement évidentes (annexe B.3). La dimension D6 valide la cohérence des métadonnées (labels, méthodes de génération, split assignments).

2.3 Sélection des quatre détecteurs

Les quatre détecteurs de deepfakes GAN-trained ont fait l'objet d'un processus de sélection rigoureux afin de maximiser la complémentarité des approches de détection. Ce critère est crucial pour que les approches d'ensemble puissent exploiter des signaux distincts plutôt que redondants. Les détecteurs sont sélectionnés selon la taxonomie de DeepfakeBench (SCLBD, 2026) qui distingue trois types : les détecteurs naïfs (CNN standard sans algorithme de détection spécialisé), les détecteurs spatiaux (CNN avec algorithmes avancés comme l'apprentissage contrastif), et les détecteurs fréquentiels (analyse du domaine de Fourier), comme illustré par la Figure 4. Cette diversité est complétée par la variation des approches d'apprentissage (standard vs contrastif) et des backbones (léger vs profond). Les architectures sélectionnées et leurs poids respectifs entraînés sur FaceForensics++ c23 ont été récupérés au sein du benchmark de référence (SCLBD, 2026).

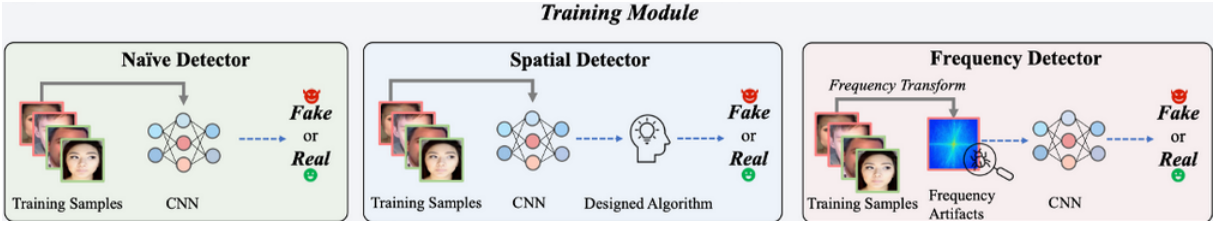

Figure 4 : Illustration des différents types de détecteurs (SCLBD, 2026)

2.3.1 Meso4 : baseline de référence historique

Meso4 représente le détecteur le plus simple de l'ensemble d'un point de vue architectural et conceptuel. Introduit par Afchar et al. en 2018, Meso4 repose sur une architecture MesoNet utilisant seulement quatre couches convolutionnelles suivies de couches fully-connected totalisant moins d'un million de paramètres. Cette conception minimaliste fait de Meso4 un modèle rapide à inférer mais limité en capacité de détection : il capture les artefacts spatiaux de bas niveau, se voyant ainsi souvent dépassés par les deepfakes évolués et raffinés (Afchar et al., 2018). Malgré des performances modestes (AUC de 0.608 sur FF++ c23), le choix de sélectionner Meso4 se justifie par le caractère historique et populaire du modèle. Il nous permettra également de contextualiser les gains liés aux architectures profondes (SCLBD, 2026).

2.3.2 XceptionNet : référence académique standardisée

XceptionNet représente le détecteur spatial de référence au sein de la sphère de détection. L'architecture Xception fut introduite par Chollet en 2017. Ce modèle repose sur le principe

des convolutions séparables en profondeur, soit une variante des convolutions standards qui réduit la complexité opérationnelle de l'opération par la factorisation des opérations spatiales et canaux, tout en préservant la capacité de détection (Chollet, 2017). XceptionNet constitue le backbone central du benchmark FF++ et obtient un score AUC de 0.964 sur FF++ c23, ce qui en fait une référence au sein de la littérature scientifique (SCLBD, 2026).

2.3.3 UCF : détecteur contrastif avancé

UCF (Universal Contrastive Forensics) représente l'approche la plus contrastive de la sélection. Ce modèle, basé sur un backbone Xception, repose sur le principe d'apprentissage via une fonction de perte contrastive qui encourage le modèle à rapprocher les représentations d'images d'une même classe tout en éloignant celles de l'autre classe au sein d'un espace latent (Yan et al., 2023). Ce concept permet à UCF d'apprendre directement des représentations invariantes indépendamment des particularités stylistiques spécifiques à un dataset d'entraînement, lui permettant ainsi théoriquement de mieux généraliser en dehors de sa zone d'entraînement. Cette hypothèse reste à valider au sein de notre recherche. Sur le dataset d'entraînement FF++ c23, UCF atteint d'excellentes performances avec une AUC de 0.971.

2.3.4 F3Net : exploitation du domaine fréquentiel

F3Net (Frequency-Aware Decomposition Network) est le représentant des détecteurs fréquentiels au sein de cette sélection et obtient une excellente AUC de 0.964 sur FF++ c23. Introduit par Qian et al. en 2020, F3Net repose sur une architecture combinant à la fois une branche spatiale classique (backbone Xception) et une branche fréquentielle FAD (Frequency-Aware Decomposition). La branche FAD décompose l'image en multiples bandes de fréquences via la transformée de Fourier discrète bidimensionnelle (2D-DFT), puis analyse les composantes fréquentielles invisibles dans le domaine spatial comme les contours, la texture, les artefacts de compression, la structure globale ou encore les variations lumineuses (Qian et al., 2020). Les GAN sont très exposés à ces détecteurs par les artefacts fréquentiels qu'ils génèrent.

2.3.5 Complémentarité architecturale et poids gelés

La sélection finale retenue tente de maximiser la complémentarité architecturale entre les détecteurs sur base des aspects suivants : le type de détection (naïf, spatial et fréquentiel), l'approche d'apprentissage (standard ou contrastif), et le niveau de sophistication (de Meso4

avec moins d'1M de paramètres à F3Net avec plus ou moins 26M de paramètres). Le Tableau 3 présente une synthèse des caractéristiques sur nos quatre détecteurs.

Le fait que ces détecteurs soient utilisés avec leurs poids pré-entraînés sur FF++ c23 et complètement gelés implique que toute augmentation des performances par l'un des modèles d'ensemble est exclusivement provoquée par la combinaison intelligente des détecteurs. Cela nous permet de quantifier directement le gain lié à notre question de recherche concernant la capacité des modèles d'ensemble à récupérer de la robustesse face aux modèles de diffusion en exploitant uniquement des détecteurs GAN existants.

Détecteur	Type	Backbone	Paramètres	Apprentissage	AUC FF++ c23	Rôle dans l'ensemble
Meso4	Naïf	MesoNet	<1M	Standard	0,608	Baseline historique
XceptionNet	Naïf	Xception	~23M	Standard	0,964	Référence spatiale
UCF	Spatial	Xception	~23M	Contrastif	0,971	Généralisation contrastive
F3Net	Fréquentiel	Xception + FAD	~26M	Standard	0,964	Analyse fréquentielle

Tableau 3 : Comparaison des quatre détecteurs GAN-trained sélectionnés

2.4 Stratégies d'ensemble : trois scénarios complémentaires

L'approche d'ensemble constitue la contribution principale apportée par cette recherche. Cette contribution émane du constat que (ré)entraîner un détecteur de deepfakes est extrêmement coûteux et énergivore. Ainsi, cette stratégie explore la possibilité de récupérer de la robustesse en combinant intelligemment des détecteurs entraînés sur un dataset issu de GAN mais appliqués sur des DM. Les trois scénarios élaborés sont ainsi détaillés dans ce chapitre.

2.4.1 Scénario A : Moyenne arithmétique simple

Le scénario A implémente la stratégie théorique la plus simple et intuitive : la moyenne arithmétique simple. Pour chaque image, les quatre détecteurs produisent chacun une probabilité P(FAKE) comprise entre 0 et 1. Le score d'ensemble est calculé comme la moyenne arithmétique simple de ces quatre probabilités (pondération de 25% par détecteur) :

$$Score_A = \frac{\left(P_{Meso4} + P_{XceptionNet} + P_{UCF} + P_{F3Net}\right)}{4}$$

Ce scénario ne nécessite aucun entraînement, aucun paramètre à calibrer, et peut être déployé instantanément dès que les quatre détecteurs ont produit leurs prédictions.

2.4.2 Scénario B : Moyenne pondérée par accuracy

Le Scénario B vient améliorer le principe de la moyenne arithmétique simple (scénario A) en adaptant la pondération de chaque détecteur sur base de ses performances respectives. Ces poids sont dès lors proportionnels aux performances de précision mesurées sur le Training Set. Concrètement, pour chaque détecteur i, son accuracy sur l'ensemble d'entraînement Acc_i est calculée, puis un poids normalisé w_i est dérivé tel que la somme des poids égale 1 :

$$w_i = \frac{Acc_i}{(Acc_{Meso4} + Acc_{XceptionNet} + Acc_{UCF} + Acc_{F3Net})}$$

Le score d'ensemble est ensuite calculé comme la moyenne pondérée des probabilités :

$$Score_B = w_{Meso4} \times P_{Meso4} + w_{XceptionNet} \times P_{XceptionNet} + w_{UCF} \times P_{UCF} + w_{F3Net} \times P_{F3Net}$$

Cette approche a pour avantage de donner plus d'importance aux détecteurs performants, ce qui doit à priori améliorer les performances. Le calcul des poids ne nécessite qu'une seule passe sur le Training Set rendant cette calibration extrêmement légère comparée à un réentraînement complet.

2.4.3 Scénario C : Méta-learner logistique

Le scénario C est la stratégie d'ensemble la plus sophistiquée dans notre recherche. Elle consiste en un méta-learner entraîné par régression logistique sur les scores des détecteurs dans le Training Set. Techniquement, le méta-learner apprend des coefficients β optimaux via la minimisation d'une fonction de perte log-loss :

$$Score_C = \sigma(\beta_0 + \beta_{Meso4} \times P_{Meso4} + \beta_{XceptionNet} \times P_{XceptionNet} + \beta_{UCF} \times P_{UCF} + \beta_{F3Net} \times P_{F3Net})$$

où σ représente la fonction sigmoïde transformant le score linéaire en probabilité, et les coefficients β sont appris par descente de gradient stochastique avec régularisation L2 pour prévenir le surapprentissage.

Contrairement aux scénarios A et B, le méta-learner peut apprendre des coefficients négatifs si un détecteur produit un signal inversement corrélé avec la vérité terrain, exploitant ainsi des

patterns de complémentarité plus subtils. Son entraînement se limite strictement au Training Set et représente une charge computationnelle négligeable contrairement à un réentraînement complet.

2.5 Métriques d'évaluation

La sélection des métriques d'évaluation a fait l'objet d'une réflexion rigoureuse afin d'interpréter les résultats de la façon la plus optimale. Ainsi, nos résultats seront considérés à travers quatre métriques complémentaires : l'AUC-ROC, l'Accuracy, le score F1 et l'Equal Error Rate (EER).

2.5.1 AUC-ROC : métrique principale

L'Area Under the Receiver Operating Characteristic Curve (AUC-ROC) constitue la métrique principale de cette recherche. Dans notre cas, cette métrique mesure la probabilité que le modèle, s'il reçoit une image réelle et une image synthétique choisies aléatoirement, attribue un score plus élevé à l'image synthétique qu'à l'image réelle. Autrement dit :

$$AUC = P\big(score(FAKE) > score(REAL)\big)$$

Une AUC de 1.0 signifie une discrimination parfaite (toutes les images FAKE reçoivent un score supérieur à toutes les images REAL), tandis qu'une AUC de 0.5 correspond à une performance aléatoire (le modèle ne performe pas mieux qu'un tirage totalement aléatoire).

Le choix de l'AUC comme métrique principale est justifié pour plusieurs raisons. Premièrement, contrairement à l'Accuracy ou le F1-score, l'AUC évalue tous les seuils de décision possibles simultanément fournissant ainsi une mesure de performance indépendante des choix de calibration. Deuxièmement, l'AUC reste stable même lorsque le ratio FAKE/REAL varie. Ce point est particulièrement crucial pour notre analyse par méthode de diffusion où les effectifs varient fortement. Enfin, l'AUC-ROC constitue la métrique standard utilisée au sein de la littérature scientifique sur la détection de deepfakes.

2.5.2 Accuracy, F1-Score et Equal Error Rate : métriques complémentaires

L'Accuracy mesure la proportion de prédictions correctes parmi l'ensemble des prédictions, en appliquant un seuil de décision fixé à 0.5 :

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

où TP = True Positives, TN = True Negatives, FP = False Positives, et FN = False Negatives.

L'Accuracy est la mesure la plus intuitive mais présente une faiblesse critique en cas de déséquilibre des classes, ce qui sera notre cas lors de l'analyse par méthode de diffusion. En effet, si une classe est majoritaire dans notre échantillon de données, un classifieur prédisant systématiquement cette classe aura une Accuracy élevée sans pour autant être performante d'un point de vue discrimination. Cette métrique est performante sur les résultats liés à notre dataset entier étant donné l'équilibre parfait entre la classe FAKE et la classe REAL.

Le F1-Score représente la moyenne harmonique de la précision et du rappel, pondérant équitablement les erreurs de type I (faux positifs) et de type II (faux négatifs) :

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

En pénalisant les valeurs extrêmes, il garantit qu'un score élevé n'est atteint que lorsque la précision et le rappel sont tous deux raisonnablement élevés. Cette métrique est particulièrement pertinente dans un contexte où les deux types d'erreurs ont des impacts similaires : un faux positif est aussi grave qu'un faux négatif, ce qui est notre cas.

Enfin, l'Equal Error Rate (EER) identifie le seuil de décision où le taux de faux positifs égale le taux de faux négatifs, fournissant ainsi un point d'équilibre opérationnel où les deux types d'erreurs sont également probables. Un EER faible indique qu'il existe un seuil permettant de minimiser simultanément les deux types d'erreurs. L'EER complète l'AUC en fournissant un point de référence unique sur la courbe ROC, permettant de comparer directement les modèles à un régime opérationnel spécifique.

2.6 Protocole expérimental complet

Le protocole expérimental de cette recherche s'organise suivant un flux en 6 phases. Chaque phase est rattachée à un notebook spécifique. Ces notebooks sont disponibles au sein du projet GitHub relatif à cette étude (Annexe A :).

La première phase consiste en l'extraction et la préparation des données. Pour ce faire, les images ont été extraites de DeepfakeBench v2 selon le protocole vidéo level défini. Les images ont été classifiées et associées à leurs métadonnées. La seconde phase consiste en l'audit des données défini en section 2.2.5. La troisième phase consiste en l'application du split-vidéo level en trois ensembles ainsi qu'à l'isolement du Validation Set afin de ne pas l'ouvrir par erreur. La quatrième phase contient le processus d'inférence des quatre détecteurs. C'est-à-dire l'association des poids aux architectures et les premières prédictions des quatre détecteurs indépendants sur le Test Set. La cinquième phase consiste en l'élaboration et l'évaluation des scénarios d'apprentissage d'ensemble, comprenant la phase de calibration des poids pour le scénario B et l'entraînement du méta-learner pour le scénario C. Les poids sont enregistrés et appliqués tels quels sur les Test Set et puis le Validation Set déverrouillé afin d'obtenir les résultats finaux. La sixième phase concerne l'élaboration de plusieurs analyses complémentaires nous permettant de répondre à nos questions de recherche. La première décompose les performances des détecteurs et ensembles par méthode de diffusion. La deuxième permet de calculer les corrélations d'erreurs inter-modèles via le coefficient phi (corrélation d'erreurs binaires). La troisième analyse deux scénarios d'ablation sur le méta-learner (Annexe F).

Le protocole complet garantit une séparation stricte entre les phases de calibration (exclusivement sur Training Set) et d'évaluation (Test Set puis Validation Set). Aucune décision méthodologique n'a été informée par l'observation des performances sur le Validation Set, respectant ainsi le principe du coffre-fort méthodologique et garantissant que les métriques finales reflètent une véritable généralisation. De plus, le biais d'HARKing (Hypothesizing After Results are Known) sera rigoureusement évité. Celui-ci consiste à dire qu'une fois les résultats connus, procéder à une modification des processus d'évaluation représenterait un biais méthodologique où les hypothèses sont ajustées rétroactivement pour s'aligner sur les résultats observés. Cette rigueur méthodologique est d'autant plus importante que notre travail est explicitement cadré comme une évaluation de modèles existants, et non comme une optimisation vers le meilleur modèle possible (Wilson, s. d.).

Chapitre 3 : Résultats

Ce chapitre présente les résultats empiriques obtenus au terme des expérimentations réalisées. Ces résultats sont présentés à travers les quatre métriques d'évaluation définies en section 2.5. De plus, ceux-ci sont organisés autour de quatre axes principaux répondant directement aux contributions scientifiques liées à nos questions de recherche : la quantification de l'obsolescence des détecteurs individuels, l'évaluation de la robustesse récupérée à travers les modèles d'ensemble, l'analyse du gradient systématique de détectabilité vis-à-vis des architectures de détection et l'étude de complémentarité entre détecteurs. L'ensemble de ces résultats proviennent du Validation Set (2811 images).

Un premier résultat méthodologique crucial concerne la stabilité des performances de ces résultats par rapport à ceux issus du Test Set. L'annexe C.1 présente les écarts AUC entre ces deux ensembles. Tous les détecteurs et scénarios testés présentent des écarts AUC inférieurs à 1.5 points de pourcentage, ce qui valide la robustesse de notre processus expérimental et l'absence de sur-apprentissage.

Il est également important de noter que ces résultats appartiennent directement au dataset sur lequel ils sont évalués. Le changement de la pondération de représentativité des modèles de deepfakes pourrait, par exemple, influencer les performances et donc les résultats obtenus, au même titre qu'une variation des paramètres d'entrées des détecteurs ou encore le choix même de ceux-ci. Ces résultats ne sont donc pas directement comparables aux résultats issus d'études ne travaillant pas sur le même dataset. Cependant, ils permettent de dégager des tendances et ainsi tirer des conclusions qui contribuent directement à la recherche scientifique, bien qu'une contextualisation de celles-ci soit indispensable.

3.1 Performances des détecteurs individuels : quantification de l'obsolescence

La première phase d'inférence nous permet d'obtenir des résultats spécifiques à chaque détecteur indépendamment des autres. Ces résultats quantifient la dégradation des performances des détecteurs GAN-trained sur notre jeu de données constitué de DM. Le Tableau 4 présente les métriques de performance associées à chaque détecteur sur le Validation Set (VS), permettant de mesurer l'obsolescence absolue. Ces résultats ont fait l'objet d'un test statistique par bootstrap (1000 itérations) qui atteste de la significativité des résultats (Annexe E.1).

Modèle	AUC	Accuracy	F1-Score	EER	P(FAKE) minimum	P(FAKE) moyen	P(FAKE) maximum
XceptionNet	0.7696	0.6403	0.5670	0.2882	0.0004	0.3943	0.9999
UCF	0.7669	0.6624	0.5555	0.3084	0.0011	0.2780	0.9998
F3Net	0.7513	0.6567	0.5442	0.3205	0.0003	0.2909	1.000
Meso4	0.5017	0.5005	0.6670	0.5052	0.4694	0.7134	0.9620

Tableau 4 : Résultats des détecteurs individuels sur le VS

À première vue, les résultats révèlent la présence d'un trio de tête aux performances internes plus ou moins similaires avec XceptionNet, UCF et F3Net. Meso4, quant à lui, dispose de résultats fortement éloignés de ces trois modèles et subit un effondrement complet face au jeu de données DM testés. Celui-ci dispose d'une AUC de 0.502 sur le Validation Set, soit une performance que l'on ne pourrait statistiquement pas distinguer d'une évaluation totalement aléatoire. La mesure d'Accuracy confirme ce constat : Meso4 dispose de 50.05% de précision sur notre dataset, sachant que celui-ci est parfaitement équilibré entre images réelles et synthétiques. En réalité, ces performances médiocres s'expliquent par le fait que Meso4 attribue presque systématiquement les images à la classe FAKE : la moyenne des scores P(FAKE) de Meso4 est de 0.7134 avec un minimum de 0.4694. Ce résultat d'obsolescence totale était attendu compte tenu de l'architecture minimaliste de Meso4 et de ses performances modestes sur les GAN (AUC de 0.61 en intra-domaine FF++ c23) (SCLBD, 2026).

Par ailleurs, Meso4 obtient un F1-score de 0.667 qui peut sembler contre intuitif compte tenu de son score AUC et son Accuracy. Cette anomalie est en réalité provoquée par son biais systématique de classification qui lui donne un score Recall très élevé sur la classe FAKE au prix d'une faible précision. Le F1-Score, étant la moyenne harmonique de Precision et Recall, est artificiellement gonflé par le Recall élevé. Ce cas illustre la faiblesse de la métrique F1-Score lorsque le modèle présente un biais systématique.

En revanche, XceptionNet, UCF et F3Net manifestent une dégradation graduée mais beaucoup plus contrôlée avec des scores AUC respectivement de 0.7696, 0.7669 et 0.7513. Autrement dit, pour XceptionNet, une paire d'images REAL/FAKE sélectionnée aléatoirement a 76.96% de chance de donner un score plus élevé à l'image synthétique, soit une chute de 19 points sur notre dataset par rapport à ses performances sur son dataset GAN d'entraînement (AUC de 0.96). Cette dégradation confirme la perte de performance énoncée dans la littérature, bien qu'elle n'annihile pas totalement la capacité du détecteur à capter des artefacts discriminants.

Bien que le détecteur UCF soit le meilleur en termes d'Accuracy avec 66.24% de prédictions correctes, il demeure légèrement inférieur à XceptionNet sur notre métrique AUC (0.7669). Ce constat est étonnant étant donné que son architecture contrastive est théoriquement plus généralisable. Ce résultat suggère que la rupture architecturale des DM face aux GAN dépasse les capacités de l'apprentissage contrastif à généraliser.

Enfin, F3Net atteint une AUC de 0.7513, légèrement inférieure aux deux détecteurs précédents. Ce résultat est contre-intuitif puisque F3Net est censé exploiter les artefacts fréquentiels distincts entre GAN et DM comme documentés dans la littérature. Ce constat s'explique plausiblement par le fait que les artefacts spectraux des DM sont suffisamment subtils pour que même la branche FAD ne parvienne pas à les capturer. Il dépasse tout de même XceptionNet avec 65.67% de précision.

Ces résultats laissent toutefois paraître une forme de décalage entre les scores AUC et Accuracy. Cela est lié à un problème de calibrage et s'explique à l'aide du Tableau 4, présentant les scores P(FAKE) minimum, moyen et maximum de chaque détecteur. La moyenne P(FAKE) des détecteurs compétents (XceptionNet : 0.39, UCF : 0.28, F3Net : 0.29) est systématiquement inférieure au seuil de décision standard de 0.5, révélant une sous-confiance caractéristique des détecteurs confrontés à des artefacts hors-distribution. Cette compression des scores vers le bas explique mécaniquement pourquoi l'AUC (qui mesure l'ordre relatif des prédictions) demeure substantielle (0.75-0.77) tandis que l'Accuracy (qui évalue les décisions binaires au seuil 0.5) chute significativement (0.64-0.66). Les F1-Score d'environ 0.55 indiquent que les détecteurs détectent correctement certaines FAKE (précision préservée) mais en ratent beaucoup (rappel faible), conséquence directe de la sous-confiance observée dans le Tableau 4. Les matrices de confusion de chaque détecteur confirment également cette surreprésentation des faux négatifs (annexe C.3 Matrices de confusion).

Toutefois, comme détaillé dans les limites (section 4.3), une partie de la dégradation des performances par rapport à leur environnement d’entraînement peut également être liée directement à la compression facteur 23 de FF++. Il convient de rester vigilant sur l’interprétation des causes provoquant cette chute de performance.

3.2 Performances des ensembles : récupération pragmatique de robustesse

Une des contributions scientifiques de ce travail évalue la capacité des approches d'ensemble à récupérer de la robustesse sans réentraînement des détecteurs de base. Le Tableau 5 présente les résultats obtenus par nos trois scénarios d'agrégation sur le Validation Set, permettant de mesurer la robustesse apportée par rapport à nos détecteurs individuels.

Modèle / Scénario	AUC	Accuracy	F1-Score	EER	Moyenne P(FAKE)	Minimum P(FAKE)	Maximum P(FAKE)
Meso4	0,5017	0,5005	0,6670	0,5052	0.7134	0,4694	0,962
XceptionNet	0,7696	0,6403	0,5670	0,2882	0.3943	0,0004	0.9999
UCF	0,7669	0,6624	0,5555	0,3084	0.2780	0,0011	0.9998
F3Net	0,7513	0,6567	0,5442	0,3205	0.2909	0,0003	1.000
Scénario A	0,7800	0,6816	0,6556	0,2967	0,4165	0,1295	0,9775
Scénario B	0,7826	0,6834	0,6048	0,2960	0,3964	0,1054	0,9816
Scénario C	0,7884	0,6958	0,6638	0,2878	0,4908	0,1536	0,9576

Tableau 5 : Résultats des détecteurs et scénarios sur le VS

Les trois scénarios d'ensemble élaborés fournissent de meilleurs résultats que les détecteurs individuels, validant ainsi l'hypothèse centrale selon laquelle la combinaison de signaux complémentaires améliore la robustesse même sans réentraînement. Le scénario A (moyenne arithmétique simple) atteint une AUC de 0.780, soit un gain de 1.04 points de pourcentage (pp) par rapport au détecteur individuel le plus performant (XceptionNet). Le scénario B, venant complexifier le scénario A par le principe de moyenne pondérée, gagne seulement 0.26 pp sur celui-ci et atteint un taux de précision de 68.34%. Enfin, le scénario C contenant le méta-learner obtient les résultats les plus élevés de nos expérimentations avec une AUC de 0.7884 et une précision de 69.58%, soit respectivement des gains de 1.88 pp et 5.55% par rapport à XceptionNet. Une analyse bootstrap (1 000 itérations, rééchantillonnage stratifié) confirme la significativité statistique de ce gain : l'IC 95% du ΔAUC est [0.0104, 0.0272], excluant la valeur zéro ($p < 0.001$) (voir Annexe E).

L'analyse des métriques révèle que les scores Accuracy et F1-Score augmentent de façon plus prononcée que le score AUC pour les scénarios A et C. Ce constat s'explique par la capacité des modèles d'ensemble à mieux calibrer les scores de confiance, rapprochant la moyenne P(FAKE)

du seuil de décision de 0.5. En effet, le méta-learner atteint une quasi-parfaite centration à 0.4908.

Dans le cas du Scénario A, la hausse du P(FAKE) moyen est justifiée par l'effet de Meso4 qui vient élever la moyenne par son propre P(FAKE) moyen de 0.7134, portant le score d'ensemble à 0.4165. Bien que toujours présent, cet effet est atténué dans le Scénario B, étant donné le poids normalisé de 20.34 % attribué à Meso4, ramenant la moyenne P(FAKE) à 0.3964. Malgré un F1-Score bien inférieur à celui du Scénario A (0.6048 contre 0.6556), les performances d'Accuracy restent supérieures de 0.18 pp, attestant l'efficacité de la pondération ajustée.

Le Scénario C se distingue par une amélioration systématique de toutes les métriques : AUC de 0.7884, Accuracy de 0.6958 et F1-Score de 0.6638. Contrairement aux Scénarios A et B qui attribuent des poids positifs fixes à tous les détecteurs, le méta-learner optimise une régression logistique dont les coefficients β reflètent la contribution nette de chaque détecteur, pouvant aller dans les valeurs négatives comme pour Meso4 : $\beta_{Meso4} = -1.99, \beta_{XceptionNet} = 1.79, \beta_{UCF} = 1.19, \beta_{F3Net} = 1.18$.

Le méta-learner attribue un poids négatif mais conséquent à Meso4 signifiant que le méta-learner interprète une faible P(FAKE) lorsque le détecteur produit un P(FAKE) élevé. C'est précisément ce mécanisme d'inversion qui hausse la moyenne P(FAKE) à 0.4908, offrant au scénario C la meilleure centration de notre étude.

La taille relativement modeste du Training Set (1402 images) soulève des questions sur la stabilité de ces coefficients appris, notamment le β fortement négatif de Meso4 (-1.99). Afin de s'assurer que le mécanisme d'inversion observé n'était pas simplement dû au hasard échantillonnal, une analyse de sensibilité par bootstrap a été réalisée sur le Training Set (1 000 itérations, rééchantillonnage stratifié). Les résultats ne laissent aucune ambiguïté : sur l'ensemble des itérations, le coefficient β_{Meso4} oscille autour d'une moyenne de -2.01 (IC 95% : [-2.94, -1.09], σ = 0.47), et reste systématiquement négatif dans 100% des cas (t = -133.74, p < 0.001). L'intervalle de confiance n'inclut jamais la valeur zéro, ce qui exclut toute interprétation du phénomène comme un bruit statistique. Le méta-learner exploite donc de manière reproductible l'inversion du signal de Meso4 pour améliorer ses prédictions. L’analyse détaillée se trouve sous l’annexe D.

La Figure 5 confirme cette interprétation : le gain du scénario C sur XceptionNet est massivement concentré dans la région FPR ∈ [0.08, 0.22]. En effet, XceptionNet seul souffre d'une distribution P(FAKE) moyenne de 0.3927 signifiant qu'il est sous-calibré vers REAL. En inversant le signal de Meso4 pour corriger le biais global, le méta-learner remonte son FPR précisément dans cette zone critique.

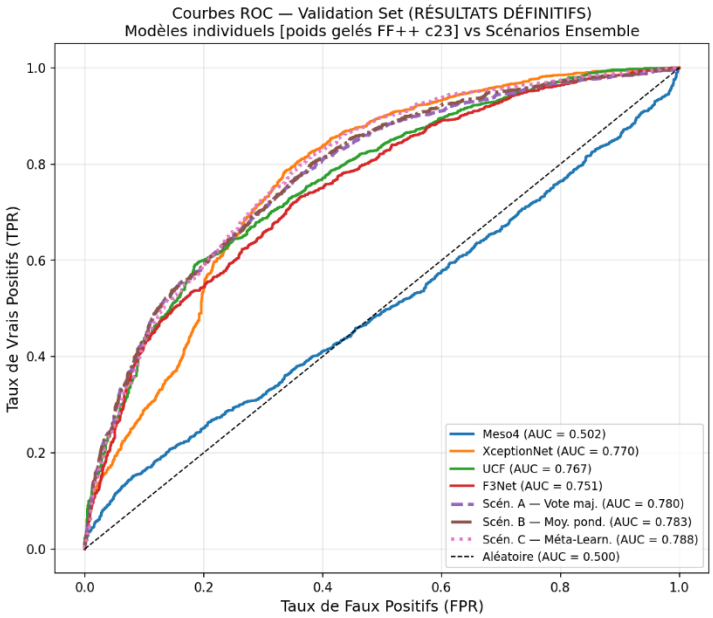

Figure 5 : Courbes ROC des détecteurs et scénarios sur le VS

Ces résultats démontrent la capacité d'une approche d'ensemble légère à combiner des signaux complémentaires, mais également à corriger les biais de calibration liés à l'évaluation hors domaine. Toutefois, l'interprétation du signal inverse de Meso4 doit être nuancée. L'analyse d'ablation présentée en Annexe F démontre qu'en remplaçant Meso4 par un classifieur trivial constant, le méta-learner obtient des résultats quasiment identiques (Δ 0.08 pp). Cela indique que Meso4 n'apporte aucune valeur ajoutée nette au méta-learner : l'effet de recalibration lié à sa moyenne P(FAKE) élevée est entièrement reproduit par l'intercept du méta-learner dès que Meso4 est retiré. Sa contribution est donc mécanique et substituable, et pas informationnelle.

Cependant, le gain absolu du méta-learner reste limité. 1.8 points d'AUC représente une amélioration mesurable, soit une réduction de ~8 % de l'erreur résiduelle (1 − AUC), mais cela reste largement insuffisant pour atteindre les performances des détecteurs GAN-trained sur leur ensemble de référence (XceptionNet atteignait 0.96 d'AUC).

3.3 Gradient de détectabilité par architecture de diffusion

Une autre contribution scientifique de cette recherche teste l'hypothèse selon laquelle la distance architecturale des modèles de diffusion serait directement corrélée avec un gradient systématique de détectabilité. Les Tableau 6 et Tableau 7 décomposent ainsi les résultats obtenus précédemment par modèles de diffusion afin de quantifier empiriquement ce gradient. Le tableau reprenant toutes les métriques d'évaluation se situe sous l'annexe C.2.

Ces résultats sont obtenus à partir des évaluations sur le Validation Set, disposant de 1404 images réelles et respectivement 144, 639, 104, 520 deepfakes pour DiT, MidJourney, SiT et ddim.

Modèle / Scénario	DiT	MidJourney	SiT	ddim	AUC
Meso4	0.4820	0.4724	0.4886	0.5457	0,5017
XceptionNet	0.6299	0.7898	0.7171	0.7939	0,7696
UCF	0.5885	0.7446	0.7127	0.8547	0,7669
F3Net	0.5965	0.7951	0.6855	0.7536	0,7513
Scénario C	0.6180	0.8136	0.7296	0.8165	0,7884

Tableau 6 : Résultats AUC par DM et par détecteur/scénario sur le VS

Modèle / Scénario	DiT	MidJourney	SiT	ddim	AUC
Meso4	0.1703	0.4761	0.1291	0.4257	0,6670
XceptionNet	0.1737	0.4803	0.1951	0.5485	0,5670
UCF	0.1212	0.4779	0.2164	0.6326	0,5555
F3Net	0.1216	0.6060	0.2038	0.4594	0,5442
Scénario C	0.1983	0.6011	0.2000	0.6048	0,6638

Tableau 7 : Résultats F1-Score par DM et par détecteur/scénario sur le VS

Les résultats confirment l'hypothèse énoncée. Pour XceptionNet et UCF, les quatre méthodes se classent selon une hiérarchie cohérente du plus au moins détectable : ddim, MidJourney, SiT, DiT. F3Net constitue une exception locale : il inverse l'ordre entre ddim et MidJourney (AUC 0.795 vs 0.754), vraisemblablement parce que sa branche FAD capte des artefacts fréquentiels plus saillants dans les LDM. Cette exception ne remet pas en cause le gradient inter-famille, qui reste structurellement valide : les architectures convolutionnelles (ddim, MidJourney) demeurent systématiquement plus détectables que les Transformers (SiT, DiT) pour les trois détecteurs compétents.

Le modèle ddim se révèle comme étant le plus facilement détectable par les quatre méthodes. UCF est le détecteur le plus efficace avec un score AUC de 0.85 et un F1-Score de 0.63. XceptionNet et F3Net obtiennent des scores AUC respectifs de 0.79 et 0.75. Ces résultats s'expliquent par le fait que ddim, bien qu'étant un modèle de diffusion, utilise une architecture convolutionnelle similaire à celle des GAN sur lesquels les détecteurs ont été entraînés. Bien que le processus de génération soit différent fondamentalement (débruitage itératif vs génération directe), les artefacts produits par les convolutions locales restent partiellement connus des détecteurs GAN-trained.

MidJourney est la seconde architecture la plus détectable par XceptionNet, UCF et F3Net, atteignant des scores AUC respectifs de 0.790, 0.745 et 0.795. Le fait que F3Net surpasse XceptionNet et UCF sur MidJourney indique que ses artefacts fréquentiels sont plus marqués ou distincts que ceux de ddim, permettant à la branche FAD de F3Net de mieux les exploiter.

SiT et DiT représentent la rupture architecturale la plus profonde avec les GAN et les résultats en sont directement impactés. L'abandon des convolutions locales au profit de mécanismes d'attention globale sur des séquences de patches tokenisés génère des cohérences visuelles de long rayon face auxquelles les détecteurs GAN-trained sont structurellement étrangers. Sur DiT, même le meilleur détecteur individuel n'atteint que 0.630 d'AUC (XceptionNet), soit 13 points au-dessus du hasard. Contrairement à nos attentes, XceptionNet atteint 0.717 sur SiT et est par conséquent plus efficace que sur DiT, au même titre que UCF et F3Net. Cette inversion empirique peut s'expliquer par la très faible proportion de deepfakes DiT au sein de notre échantillon (144 images sur 2 811 au total, soit 5.1% du Validation Set, représentant seulement 9.3% du sous-échantillon DiT évalué avec les 1 404 images REAL), ce qui amplifie fortement la variance des estimations. Cette inversion entre DiT et SiT ne remet toutefois pas en cause l'hypothèse du gradient systématique de détectabilité dans sa totalité, les architectures Transformers (DiT, SiT) étant clairement et systématiquement moins détectables que les architectures à base convolutionnelle (ddim, MidJourney). Cependant, nos expérimentations ne permettent pas de valider cette hypothèse entre modèles issus de la famille Transformer (voir section 4.4).

De plus, les scores F1 ainsi que l'annexe C.4 démontrent que sur DiT et SiT, les détecteurs XceptionNet, UCF et F3Net attribuent presque systématiquement des scores P(FAKE) inférieurs à 0.5, prédisant majoritairement REAL et ne classant comme FAKE que dans des situations d'extrême confiance. Cette sous-confiance systématique explique mécaniquement comment les scores AUC demeurent légèrement supérieurs au hasard malgré une calibration défaillante.

Concernant le méta-learner (Scénario C), celui-ci affiche des performances inférieures à XceptionNet sur DiT et à UCF sur ddim, n'améliorant donc pas systématiquement les résultats sur ces méthodes de diffusion. Sur DiT, cela s'explique par la sous-représentation échantillonnaire qui induit une sous-pondération des coefficients du méta-learner optimisés sur le Training Set global. Concernant ddim, cela s'explique par le faible poids donné à UCF (β = 1.19) par le méta-learner, alors qu'UCF est justement le détecteur le plus performant sur cette

méthode spécifique. Ces résultats soulignent une limite de l'approche méta-learner : ses coefficients, optimisés sur la distribution globale du Training Set (MidJourney 45%, ddim 37%, SiT 7%, DiT 10%), reflètent les patterns majoritaires et ne peuvent s'adapter aux spécificités locales de chaque méthode sans stratification ciblée.

3.4 Complémentarité et corrélations d'erreurs entre détecteurs

L'analyse des corrélations d'erreurs inter-détecteurs permet de mesurer la complémentarité entre les détecteurs et ainsi comprendre les mécanismes sous-jacents à l'efficacité des approches d'ensemble. La Figure 6 présente la matrice de corrélation des erreurs calculée sur les erreurs binaires (correct=1, incorrect=0) des quatre détecteurs sur le VS. Appliqué à des variables binaires, ce coefficient est mathématiquement équivalent au coefficient phi, forme spécifique de la corrélation de Pearson pour variables dichotomiques. La Figure 7 présente les accords inter-modèles compétents (XceptionNet, UCF et F3Net) sur le VS.

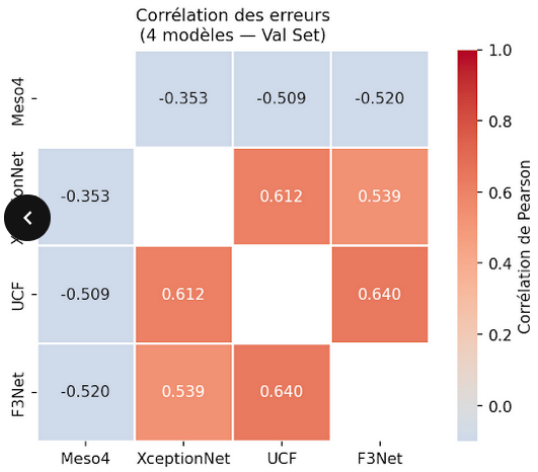

Figure 7 : Coefficient phi (corrélation d'erreurs binaires) entre détecteurs sur le VS

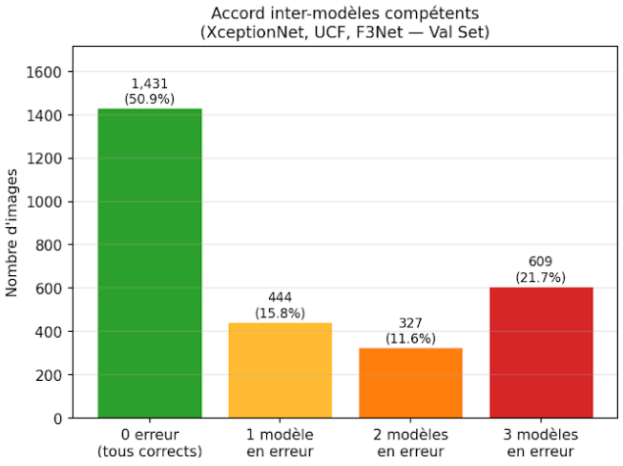

Figure 6 : Accord inter-modèles compétents sur le VS

Les résultats permettent de dissocier deux patterns fondamentaux. Tout d'abord, Meso4 présente des scores de corrélations négatifs avec les trois autres détecteurs. Cela signifie que lorsque Meso4 se trompe sur une image, les autres détecteurs ont tendance à réussir, et inversement. Ce résultat est attendu compte tenu du fait que Meso4 classe systématiquement les images dans la classe FAKE alors que les trois autres détecteurs ont tendance à classifier la majorité des images dans la classe REAL en raison du biais de calibrage identifié. Par ailleurs, cette anti-corrélation justifie la conservation méthodologique de Meso4 dans l'ensemble malgré

ses mauvaises performances : le méta-learner apprend à exploiter ce signal négatif afin d'obtenir des informations complémentaires indétectables par les autres détecteurs.

Deuxièmement, les trois détecteurs compétents présentent des corrélations d'erreurs positives modérées : 0.61 entre XceptionNet et UCF, 0.54 entre XceptionNet et F3Net, 0.64 entre UCF et F3Net. Ces résultats reflètent une redondance partielle mais pas totale entre les détecteurs, justifiant une complémentarité réelle entre détecteurs spatiaux, contrastifs et fréquentiels. La Figure 7 révèle que 27.4% des images sont en erreurs sur 1 ou 2 détecteurs parmi les trois compétents. Cet échantillon correspond au champ d'action exclusif sur lequel les modèles d'ensemble justifient leur valeur ajoutée. Cependant, cette analyse révèle également un plancher d'erreur incompressible représentant 21.7% du jeu de données sur lequel les trois détecteurs échouent simultanément.

3.5 Synthèse des résultats

Les résultats présentés dans ce chapitre apportent des réponses empiriques directes à nos deux questions de recherche structurant cette étude.

Premièrement, l'hypothèse concernant l'existence d'un gradient systématique de détectabilité corrélé à la distance architecturale des modèles de diffusion est partiellement confirmée par nos expérimentations. Tous les modèles de diffusion ne posent pas les mêmes difficultés aux détecteurs GAN-trained. La hiérarchie observée (ddim > MidJourney > SiT > DiT) se reproduit de façon stable à travers les trois détecteurs compétents, validant le gradient de détectabilité au niveau inter-famille. Cette validation reste toutefois partielle : l'inversion empirique entre DiT et SiT ne permet pas de conclure avec la même robustesse au niveau intra-famille Transformer. Au-delà de ce gradient, nos résultats révèlent la présence d'une obsolescence totale pour Meso4 (AUC 0.502) et d'une dégradation graduée pour XceptionNet, UCF et F3Net (AUC 0.751–0.770).

Deuxièmement, l'hypothèse concernant la récupération de robustesse à l'aide d'approches d'ensemble combinant des détecteurs aux signaux complémentaires a également été confirmée. Les trois scénarios évalués apportent des résultats améliorés en comparaison des détecteurs individuels. Le méta-learner du scénario C obtient les meilleures performances avec un score AUC de 0.788 sur le Validation Set, soit un gain de 1.88 point de pourcentage sur XceptionNet. Une analyse bootstrap (1 000 itérations, rééchantillonnage stratifié) confirme la significativité

statistique de ce gain : l'IC 95% du ΔAUC est [0.0104, 0.0272], excluant la valeur zéro ($p < 0.001$) (voir Annexe E). Cependant, cette récupération atteint rapidement une limite structurelle expliquée par l'existence d'un plancher d'erreur incompressible : 21.7% de l'ensemble de validation correspond à des images sur lesquelles les trois détecteurs compétents échouent simultanément.

Chapitre 4 : Discussion

Ce chapitre apporte dans un premier temps une interprétation approfondie des mécanismes sous-jacents à nos résultats. Par la suite, il positionne cette recherche vis-à-vis de la littérature existante, discute des limites méthodologiques et perspectives d'améliorations et conclut par les implications managériales.

4.1 Interprétations des mécanismes sous-jacents

Les modèles basés sur des convolutions, comme DDIM ou MidJourney, produisent une cohérence spatiale locale dont les caractéristiques rappellent les artefacts typiques des architectures GAN. C'est pourquoi les détecteurs entraînés sur des GAN parviennent à repérer ces signaux et affichent des résultats plutôt corrects (avec des AUC situées entre 0,745 et 0,855). La situation s'inverse totalement avec les architectures Transformer pures, comme DiT ou SiT. En abandonnant la convolution locale pour utiliser des mécanismes d'attention globale sur des patchs tokenisés, ces modèles génèrent des artefacts d'une tout autre nature. Logiquement, les performances des détecteurs classiques s'effondrent face à eux, tombant à des AUC comprises entre 0,589 et 0,717.

Dans ce contexte, Meso4 est complètement dépassé : il classe la quasi-totalité des images comme fausses. Son obsolescence est totale sur les modèles de diffusion, ce qui se traduit par une AUC de 0,502 — ni plus ni moins que du hasard. Ce constat est d'ailleurs validé statistiquement par l'intervalle de confiance bootstrap à 95 % [0,4816 ; 0,5223] qui englobe la valeur de référence de 0,5. Pour XceptionNet, UCF et F3Net, la baisse de régime est un peu plus progressive (AUC entre 0,751 et 0,770), mais ils accusent tous un sérieux problème de calibrage. Leurs scores P(FAKE) moyens restent bien en dessous du seuil standard (oscillant entre 0,28 et 0,39 au lieu de 0,5), ce qui révèle un net manque de confiance face à ces artefacts inédits.

Le méta-learner configuré dans le scénario C parvient à redresser la barre grâce à une double action combinée. Il corrige d'abord le biais de calibrage en inversant le signal émis par Meso4 (β = -1,99), ce qui permet de recentrer la distribution des scores P(FAKE) autour d'une valeur optimale (moyenne à 0,4908). Il convient toutefois de nuancer la portée de cette interaction, comme le met en lumière l'analyse d'ablation détaillée en Annexe F. En substituant un classifieur trivial constant à Meso4, le méta-learner conserve une AUC presque inchangée, à

peine inférieure de 0,08 point de pourcentage. Ce résultat démontre que l'apport de Meso4 ne tient pas tant à une réelle complémentarité des descripteurs qu'à un simple effet mécanique de recalibration, rendu possible par sa valeur moyenne initialement haute (0,71).

D'autre part, les corrélations d'erreurs modérées entre détecteurs compétents (0.54–0.64) indiquent que les approches spatiales et fréquentielles captent des artefacts partiellement distincts, créant un espace de complémentarité exploitable représentant 27.4% du Validation Set.

Cependant, cette récupération atteint une limite structurelle expliquée par l'existence d'un plancher d'erreur incompressible. En effet, 21.7% de l'ensemble de validation correspond à des images sur lesquelles les trois détecteurs compétents échouent simultanément. Cet échantillon représente le noyau dur d'images face auquel les méthodes d'ensemble sont totalement impuissantes.

4.2 Positionnement des résultats vis-à-vis de la littérature existante

Nos résultats interviennent dans un débat scientifique ouvert concernant la symétrie de généralisation entre les familles DM et GAN. Ricker et al. (2024) soutiennent par leurs travaux une forte asymétrie plaçant les détecteurs DM-trained bien au-dessus des détecteurs GAN-trained sur leur capacité de généralisation inter-famille (respectivement 94.5% contre 26.4% de précision). Nos taux de précision (Accuracy) vont de 64.03% à 66.24% sur nos trois détecteurs compétents et se situent substantiellement au-dessus de ce 26.4%. Cette divergence est explicable par la différence entre les détecteurs sélectionnés et le jeu de données utilisé (l'échantillon de Ricker contient Stable Diffusion XL et MidJourney v5, soit des versions potentiellement plus élaborées et difficiles à détecter). Néanmoins, nos résultats confirment la tendance générale : la dégradation graduée des détecteurs GAN-trained est substantielle et structurelle. Contrairement à Ricker et al., notre contribution réside dans la quantification précise du gradient de détectabilité. En démontrant que ddim est significativement plus détectable que DiT/SiT (écarts d'AUC jusqu'à 22 pp), nous prouvons que l'obsolescence n'est pas uniforme mais réellement corrélée à la distance architecturale entre familles de modèles de diffusion.

Ces conclusions ont été nuancées par l'étude de Choi et Woo en 2025 qui démontre que sur certaines configurations les détecteurs GAN-trained peuvent surpasser ceux DM-trained en généralisation inter-famille. Notre étude apporte une contribution supplémentaire :

XceptionNet (0.770) surpasse légèrement F3Net (0.751) globalement, malgré l'hypothèse initiale d'un avantage fréquentiel. Cela suggère que la robustesse découle autant de la profondeur et capacité des représentations que de la modalité d'analyse. Cependant, sur MidJourney, F3Net surpasse XceptionNet (0.795 vs 0.790), confirmant que l'analyse fréquentielle conserve un avantage contextuel. Ce constat nuance la contribution de Choi et Woo en démontrant qu'aucune architecture unique ne domine universellement tous les scénarios de transfert inter-famille.

À notre connaissance, cette étude constitue la première évaluation proposant l'amélioration des performances en contexte interfamilial GAN-DM par des stratégies d'ensemble à poids gelés. Les travaux antérieurs (Rana et al., 2022 ; Wahab et al., 2025 ; Naskar et al., 2024) se concentraient sur la généralisation cross-dataset intra-famille ou impliquaient un fine-tuning. Notre protocole se limite à l'apprentissage du méta-learner et simule un scénario opérationnel réaliste où les organisations ne peuvent réentraîner face à chaque nouvelle menace. Les résultats démontrent qu'un gain mesurable (1.88 pp) est accessible via cette approche légère, mais l'existence du plancher incompressible (21.7%) impose une limite structurelle. Cette limitation structurelle explique pourquoi les ensembles cross-dataset GAN-à-GAN rapportés dans la littérature atteignent des AUC > 0.85 (Rana et al., 2022), soit significativement supérieurs à notre score de 0.788 en contexte inter-familial. Dans le premier cas, les détecteurs sont confrontés à des variations superficielles d'un même type d'artefact ; dans notre cas, à l'absence totale des artefacts pour lesquels ils ont été optimisés. Cette conclusion nuancée quantifie précisément le compromis coût-performance, permettant aux organisations de décider si l'amélioration suffit ou si le réentraînement s'avère nécessaire.

4.3 Limites méthodologiques

Malgré sa rigueur méthodologique, ce travail présente certaines limitations nécessitant d'être explicitées. Premièrement, bien que le déséquilibre de représentativité entre les modèles de diffusion (MidJourney 45%, ddim 37%, DiT 10%, SiT 7%) soit justifié dans notre contexte (section 2.2.22.2.2), il biaise les métriques à la hausse. En effet, les modèles Transformers étant structurellement plus difficiles à détecter que MidJourney et ddim, un échantillon équilibré entre ces quatre modèles fournirait des performances détériorées. Les performances absolues obtenues ne sont donc pas comparables à une distribution uniforme.

Deuxièmement, les résultats sont directement conditionnés par le dataset DF40 et les configurations spécifiques des méthodes de diffusion. Des versions plus récentes (MidJourney v7, Stable Diffusion 3.5) peuvent fournir des artefacts qualitativement différents. Néanmoins, le gradient de détectabilité reflète véritablement une tendance structurelle robuste liée à la diversité architecturale des modèles de diffusion.

Troisièmement, bien que le dataset DF40 soit publié par la même équipe que DeepfakeBench (SCLBD, 2026) et donc conçu pour une intégration fluide avec le pipeline d'évaluation du benchmark, un facteur confondant résiduel persiste à cause de la compression faite sur le dataset d'entraînement FF++ c23. En effet, le preprocessing des images est sensiblement identique entre les deux datasets, minimisant le risque de shift d'acquisition. Néanmoins, la compression JPEG au facteur de qualité 23 introduit des artefacts spectraux spécifiques absents dans les images DF40. Il serait pertinent de mener ces mêmes expérimentations sur le dataset DF40 constitué d'images GAN afin de pouvoir quantifier dans quelle mesure ce shift intervient dans la chute des performances, et ainsi pouvoir isoler l'effet propre de l'architecture générative.

4.4 Perspectives d'amélioration

Quatre perspectives d'amélioration émergent de nos expérimentations. Premièrement, l'amélioration de la complémentarité des détecteurs utilisés pour les modèles d'ensemble devrait être testée par l'intégration de détecteurs spécialisés sur diffusion afin de capter plus facilement les artefacts Transformers et spécifiques aux DM. Ce modèle hybride exploiterait la complémentarité inter-famille et repousserait ainsi potentiellement le plancher au-delà de 21.7%.

Deuxièmement, l'adaptation de techniques d'apprentissage continu aux techniques d'ensemble permettrait aux coefficients du méta-learner de s'ajuster au fur et à mesure que de nouveaux échantillons sont collectés, combinant ainsi les avantages de l'ensemble (efficacité computationnelle) et du fine-tuning (adaptations aux nouveaux artefacts).

Troisièmement, l'exploration de méta-learners plus sophistiqués que la régression logistique (réseaux récurrents, mécanismes d'attention) pourrait extraire des patterns de complémentarité plus complexes, pondérant dynamiquement les détecteurs selon des caractéristiques contextuelles de chaque image.

Quatrièmement, bien que nos expérimentations permettent de valider l'hypothèse du gradient de détectabilité entre Transformer, Midjourney et ddim, cela n'est pas le cas pour les modèles au sein de la catégorie Transformer. En effet, DiT est moins facilement détectable que SiT contrairement à nos attentes sur la distance architecturale. Cette inversion s'explique par la faible proportion d'images DiT et SiT dans notre Validation Set, amplifiant fortement la variance des estimations. Une analyse approfondie avec des tailles d'échantillons augmentées mériterait d'être menée.

4.5 Implications managériales

Les résultats obtenus permettent de dégager plusieurs implications concrètes à destination des organisations confrontées à la menace des deepfakes dans leurs activités.

Premièrement, le gradient de détectabilité confirmé par nos expérimentations offre aux organisations un cadre de décision directement actionnable. Une organisation dont l'exposition aux deepfakes provient majoritairement d'architectures convolutionnelles accessibles au grand public, comme MidJourney ou ddim (vecteurs les plus répandus parmi les acteurs malveillants) peut raisonnablement s'appuyer sur des détecteurs GAN-trained existants pour maintenir un niveau de détection partiel (AUC 0.75-0.86). À l'inverse, une organisation ciblée par des deepfakes issus d'architectures Transformer (DiT, SiT) ne peut pas se contenter de cette approche : les performances chutent à des niveaux insuffisants (AUC 0.59-0.63) et le réentraînement ou le déploiement de détecteurs spécialisés devient incontournable.

Ensuite, le compromis coût-performance des approches d'ensemble est quantifié en faveur de ceux-ci. Le méta-learner apporte un gain significatif de 1.88 pp d'AUC en moins de 5 minutes contre 52 à 74 heures pour un réentraînement complet (Fontana et al., 2025). Bien que ces résultats ne soient pas précisément comparables car ils n'utilisent pas les mêmes jeux de données, la différence d'ordre de grandeur n'est pas insignifiante. En réalité, elle est décisive dans un contexte où les organisations font face à des deepfakes de plus en plus élaborés et issus de modèles de diffusion évoluant à un rythme mensuel. Ainsi, l'approche d'ensemble constitue une solution quasi-immédiate face à une nouvelle menace. Cette amélioration peut suffire lorsque le seuil de risque acceptable est proche d'être franchi ou lorsque les deepfakes ciblés proviennent majoritairement d'architectures convolutionnelles détectables.

Enfin, le plancher d'erreur incompressible de 21.7% est particulièrement contraignant pour les organisations devant minimiser le taux de faux négatifs, notamment pour les vérifications d’identité pour des accès sécurisés ou opérations bancaires. Dans ce cas spécifique, l’approche d’ensemble est insuffisante et les organisations doivent déployer les ressources nécessaires pour adapter leur système de détection (par du développement, de la recherche, et/ou du réentraînement).

En synthèse, nos résultats permettent à tout type d’organisation de positionner sa stratégie de détection sur un spectre allant de l'approche légère par ensemble — suffisante pour une exposition majoritairement convolutionnelle et un budget computationnel contraint — jusqu'au réentraînement complet, indispensable face aux architectures Transformer ou aux contextes où le coût d'un faux négatif est critique.

Chapitre 5 : Conclusion

Cette recherche avait pour objectif de quantifier l'obsolescence des détecteurs de deepfakes GAN-trained face aux modèles de diffusion et d'évaluer la capacité des approches d'ensemble à poids gelés à récupérer pragmatiquement de la robustesse sans réentraînement. Au terme de cette étude empirique menée sur 7020 images issues du dataset DF40, trois contributions scientifiques majeures peuvent être synthétisées.

Premièrement, l'existence d'un gradient systématique de détectabilité corrélé à la distance architecturale entre modèles de diffusion et GAN est confirmée par nos expérimentations, bien que partiellement au niveau intra-famille Transformer. En témoigne la hiérarchie observée par nos trois détecteurs compétents (XceptionNet, F3Net et UCF) : ddim > MidJourney > SiT > DiT. Les méthodes opérant par processus convolutionnels (ddim et MidJourney) affichent des résultats partiellement corrects (AUC 0.75-0.86), alors que les architectures Transformers pures (DiT et SiT) fournissent des performances significativement détériorées (AUC 0.59-0.72). L'inversion empirique entre ces deux dernières ne valide toutefois pas notre hypothèse entre les modèles de la famille Transformer. Cette quantification précise permet aux organisations ciblées par la menace des deepfakes de calibrer leur stratégie selon le type de génération auquel elles sont confrontées.

Deuxièmement, les stratégies d'ensemble à poids gelés permettent effectivement de récupérer partiellement de la robustesse. Le méta-learner (scénario C) atteint une AUC de 0.788, soit un gain de 1.88 pp par rapport au meilleur détecteur individuel (XceptionNet 0.770), représentant 8% de réduction de l'erreur résiduelle. Concrètement, le méta-learner atteint 69.58% de précision sur la classification binaire FAKE/REAL, soit 3.35% de plus qu'UCF, le détecteur le plus précis. Cette amélioration provient de deux mécanismes : la correction du biais de calibrage par l'inversion du signal de Meso4 ($\beta = -1.99$), et l'exploitation de la complémentarité partielle entre les détecteurs naïfs, spatiaux et fréquentiels (corrélations d'erreurs 0.54-0.64). La pertinence du méta-learner comme solution peu énergivore est également validée avec un coût computationnel de moins de 5 minutes.

Troisièmement, l'existence d'un plancher d'erreur incompressible (21.7% du Validation Set) impose une limite structurelle aux approches d'ensemble opérant exclusivement sur les scores de confiance de détecteurs gelés. Ce noyau dur d'images, correspondant vraisemblablement aux Transformers selon le gradient de détectabilité, démontre qu'au-delà d'un certain seuil

d'obsolescence, la combinaison de détecteurs défaillants ne peut compenser l'absence de représentations discriminantes appropriées. L'approche d'ensemble ne peut donc constituer qu'une solution transitoire au réentraînement lorsque le taux de représentation des deepfakes Transformers devient critique.

Ces contributions apportent des réponses empiriques précises à notre question de recherche. L'impact des modèles de diffusion sur les détecteurs GAN-trained varie de façon corrélée avec la distance architecturale entre ces modèles (sauf entre modèles intra-Transformer), une récupération partielle de robustesse est accessible via les modèles d'ensemble (gain +1.88 pp), mais cette amélioration atteint un plancher structurel (21.7% incompressible) limitant l'efficacité face aux architectures les plus éloignées des GAN. Le compromis coût-performance est également quantifié par les résultats, permettant aux organisations de trouver l'équilibre optimal entre efficacité computationnelle et robustesse de détection parmi les options à leur disposition.

Les trois principales limites méthodologiques dégagées incluent la représentation échantillonnaire déséquilibrée des modèles de génération, la conditionnalité au dataset DF40, et le shift de distribution lié à la compression c23 sur le dataset d'entraînement. Une validation sur un dataset externe indépendant et plus volumineux ajouterait une réelle plus-value pour confirmer la généralisabilité de l'approche au-delà de DF40.

Enfin, les perspectives d'amélioration s'articulent autour de quatre axes : l'intégration d'un détecteur spécialisé diffusion dans l'ensemble pour capter les artefacts Transformers manqués ; l'adaptation de techniques d'apprentissage continu permettant un ajustement incrémental sans cycles de réentraînement coûteux ; l'élaboration de méta-learners spécialisés pour extraire des patterns de complémentarité plus complexes ; et l'augmentation du nombre d'images DiT et SiT au sein de l'échantillon pour avoir des résultats statistiquement plus robustes. Ces extensions s'annoncent prometteuses dans l'optique de maximiser les performances des approches d'ensemble au sein d'une sphère qui évolue continuellement malgré son importance opérationnelle croissante.

Annexe A : Implémentation

A.1 : Code source

Le code associé à ce projet se trouve dans le répertoire GitHub : https://github.com/ducamax07/deepfake-detector-obsolescence.git

L'ensemble du pipeline expérimental a été implémenté en Python 3.10 via Google Colab afin d'avoir accès aux GPU NVIDIA nécessaires à l'inférence sur 7 020 images. Le stockage des données (DF40, poids des détecteurs, résultats intermédiaires) a été assuré par Google Drive (voir README). Par conséquent, tous les chemins de fichiers sont spécifiques à cette infrastructure (/content/drive/MyDrive/Memoire_Deepfakes/...) et ne sont pas paramétrables.

Par ailleurs, ce code a directement été élaboré et conçu exclusivement pour les expérimentations liées à cette étude et n'a pas été généralisé. L'objectif n'est pas de proposer un benchmark directement applicable mais plutôt de fournir le code source permettant de comprendre le processus opérationnel et/ou de s'en inspirer.

L'accès aux données DF40 et aux poids des détecteurs nécessite des démarches académiques indépendantes guidées par le README du projet GitHub (licence FaceForensics++, téléchargement DeepfakeBench).

A.2 Benchmark de référence

Le code du benchmark de référence à partir duquel provienne les architectures de détection, la base de données DF40 et les poids pré-entraînés se trouve sous le répertoire GitHub : https://github.com/SCLBD/DeepfakeBench.git

Annexe B : Audit de la base de données

Cette annexe fournit les graphiques, figures et tableaux pertinents à la compréhension issus de la phase d'audit sur notre ensemble de données.

B.1 : D2 Equilibre des classes

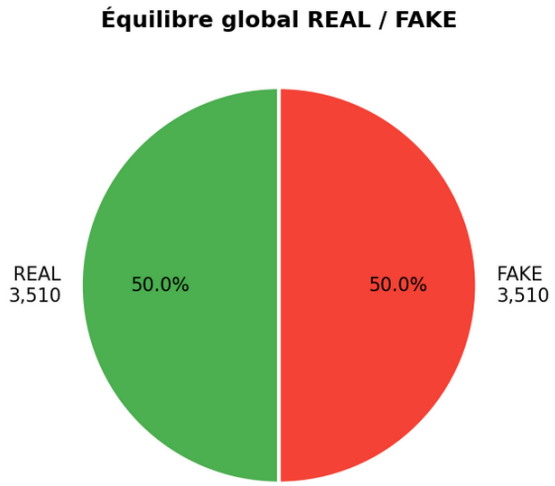

Figure 8 : Equilibre entre la classe FAKE et la classe REAL dans le dataset

B.2 : D3 Distribution statistique

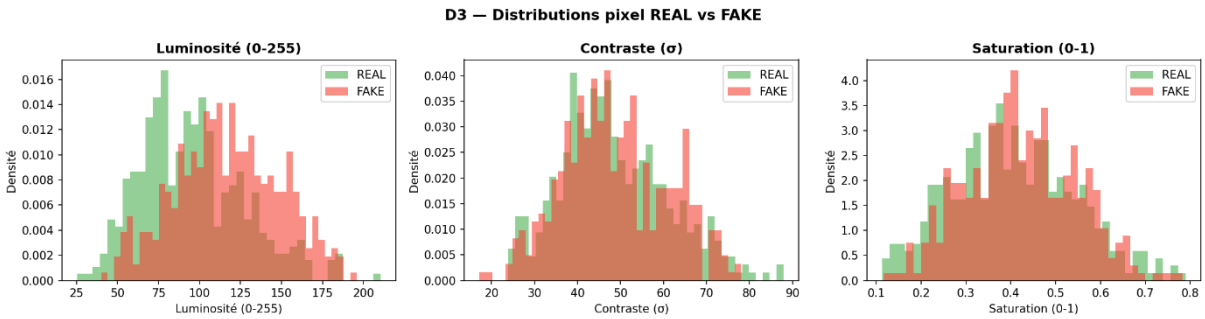

Figure 9 : Analyse de la luminosité, du contraste et de la saturation entre les images REAL et FAKE

Métrique	real_mean	real_std	fake_mean	fake_std	U_stat	p_value	significant
Luminosité	95,68	32,00	117,35	31,68	49146,00	0,00	TRUE
Contraste	48,63	12,53	48,78	12,08	78360,00	0,62	FALSE
Saturation	0,41	0,14	0,42	0,12	73033,00	0,03	TRUE

Tableau 8 : Statistiques des analyses de la luminosité, du contraste et de la saturation entre les images REAL et FAKE

B.3 : D5 Inspection visuelle

Figure 10 : Echantillon aléatoire d'images réelles

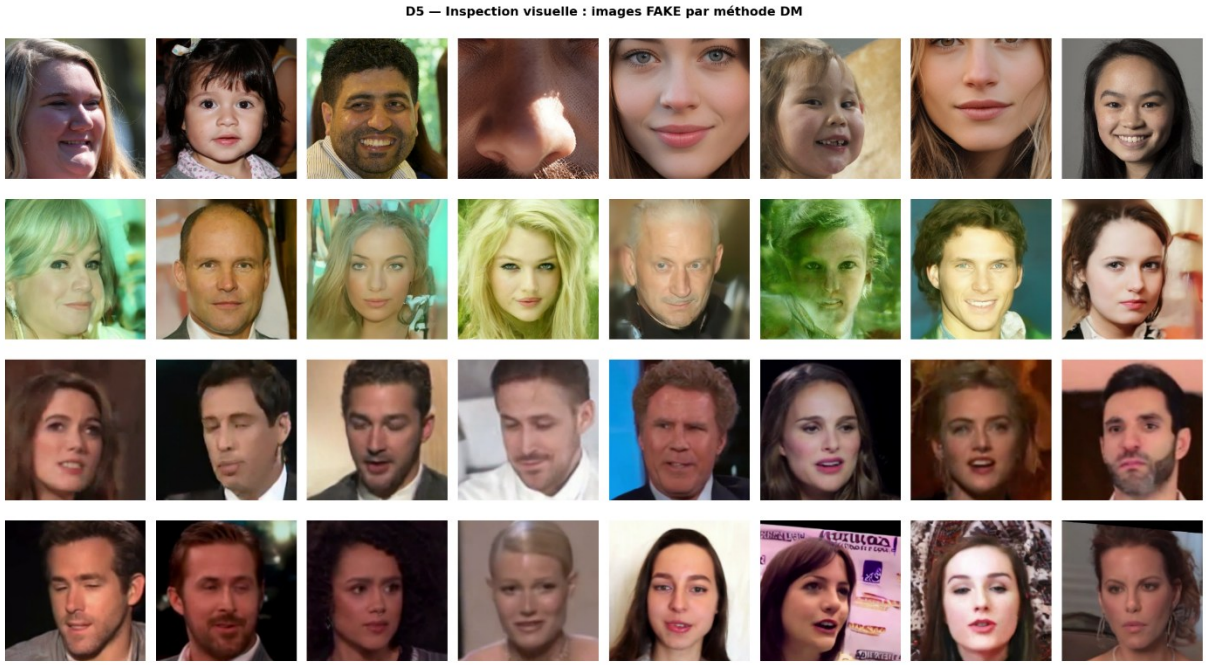

Figure 11 : Echantillon aléatoire d'images synthétiques

Annexe C : Résultats et performances

Cette annexe présente les tableaux contenant l'ensemble détaillés des résultats obtenus lors de nos expérimentations, et nécessaires à la bonne compréhension.

C.1 : Stabilité des performances entre ensemble de données

Modèle / Scénario	Catégorie	AUC	Accuracy	F1-Score	EER	Stabilité
		Δ (Val−Test)	Δ (Val−Test)	Δ (Val−Test)	Δ (Val−Test)	Max |Δ|
Meso4	Individuel	+0.0001	+0.0017	+0.0014	+0.0036	0.0036
XceptionNet	Individuel	+0.0086	+0.0129	+0.0186	−0.0121	0.0186
UCF	Individuel	+0.0146	−0.0027	−0.0061	−0.0129	0.0146
F3Net	Individuel	+0.0057	+0.0044	+0.0094	−0.0030	0.0094
Scén. A — Moy. arith.	Ensemble	+0.0097	+0.0072	+0.0093	−0.0061	0.0097
Scén. B — Moy. pond.	Ensemble	+0.0093	−0.0017	−0.0054	−0.0072	0.0093
Scén. C — Méta-Learn.	Ensemble	+0.0045	+0.0043	+0.0039	−0.0107	0.0107

Tableau 9 : Stabilité des performances Test vs Validation

C.2 : Performances par méthode de diffusion, détecteurs et scénarios d'ensemble

Méthode	Modèle	AUC	Accuracy	F1-Score	EER
DiT	Meso4	0.4820	0.0937	0.1703	0.5225
DiT	XceptionNet	0.6299	0.7603	0.1737	0.3946
DiT	UCF	0.5885	0.8314	0.1212	0.4293
DiT	F3Net	0.5965	0.8320	0.1216	0.4167
DiT	Scén. A — Moy. arith.	0.6072	0.8140	0.1579	0.4170
DiT	Scén. B — Moy. pond.	0.6101	0.8178	0.1557	0.4174
DiT	Scén. C — Méta-Learn.	0.6180	0.7494	0.1983	0.4237
MidJourney	Meso4	0.4724	0.3128	0.4761	0.5369
MidJourney	XceptionNet	0.7898	0.6970	0.4803	0.2584
MidJourney	UCF	0.7446	0.7401	0.4779	0.3314
MidJourney	F3Net	0.7951	0.7861	0.6060	0.2897
MidJourney	Scén. A — Moy. arith.	0.7997	0.7699	0.5913	0.2923
MidJourney	Scén. B — Moy. pond.	0.8033	0.7680	0.5813	0.2912
MidJourney	Scén. C — Méta-Learn.	0.8136	0.7401	0.6011	0.2778
SiT	Meso4	0.4886	0.0696	0.1291	0.5077
SiT	XceptionNet	0.7171	0.7812	0.1951	0.3346
SiT	UCF	0.7127	0.8607	0.2164	0.3374
SiT	F3Net	0.6855	0.8601	0.2038	0.3581
SiT	Scén. A — Moy. arith.	0.7187	0.8395	0.2143	0.3356
SiT	Scén. B — Moy. pond.	0.7216	0.8435	0.2133	0.3360
SiT	Scén. C — Méta-Learn.	0.7296	0.7666	0.2000	0.3392
ddim	Meso4	0.5457	0.2708	0.4257	0.4558
ddim	XceptionNet	0.7939	0.7458	0.5485	0.2885
ddim	UCF	0.8547	0.8170	0.6326	0.2188
ddim	F3Net	0.7536	0.7614	0.4594	0.3138
ddim	Scén. A — Moy. arith.	0.8157	0.7957	0.6026	0.2504
ddim	Scén. B — Moy. pond.	0.8172	0.7973	0.6012	0.2564
ddim	Scén. C — Méta-Learn.	0.8165	0.7609	0.6048	0.2598

Tableau 10 : Performances par méthode de diffusion

C.3 Matrices de confusion

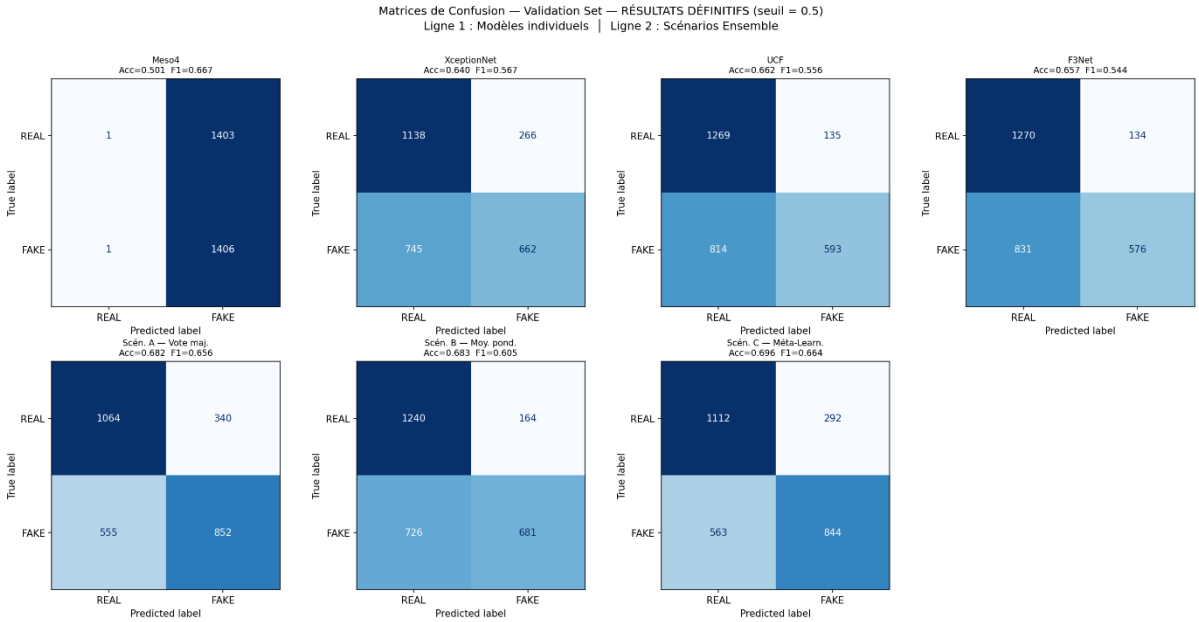

Figure 12 : Matrices de confusion des détecteurs et scénarios

C.4 Probabilités P(FAKE)

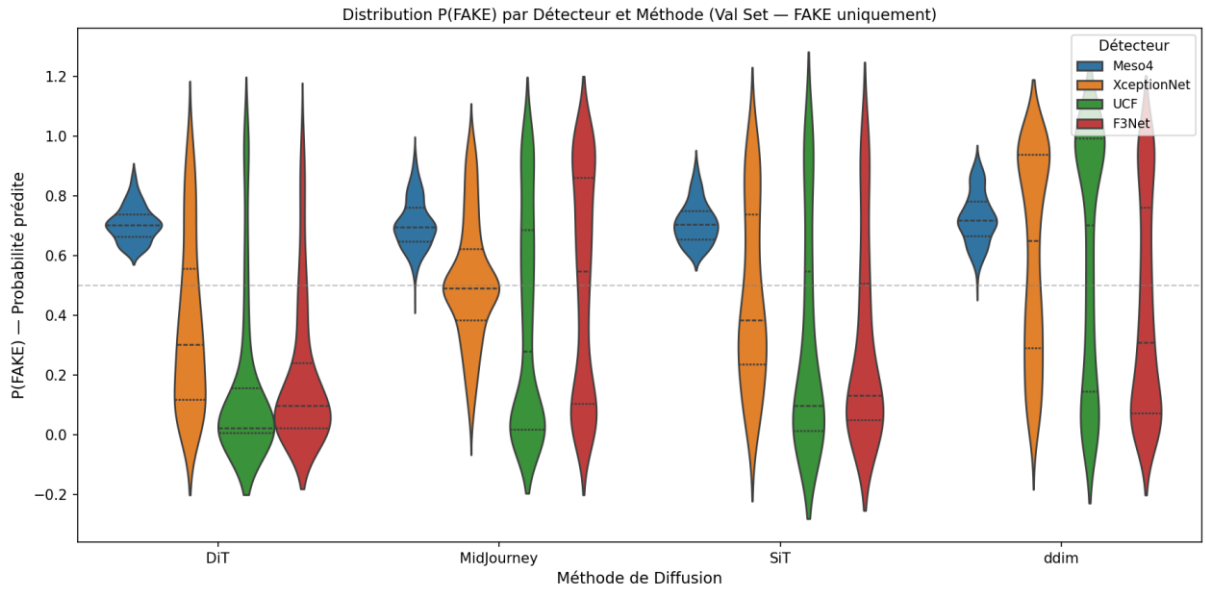

Figure 13 : Distribution P(FAKE) par détecteurs et méthodes de diffusion

Annexe D : Analyse de sensibilité du méta-learner (Scénario C)

Cette annexe valide statistiquement la robustesse du coefficient $\beta_{Meso4} \approx -1.99$ observé dans le méta-learner du Scénario C. Une analyse bootstrap (1 000 itérations, rééchantillonnage stratifié) a été conduite sur le Training Set (n = 1 402) pour confirmer que ce coefficient fortement négatif n'est pas un artefact échantillonnal. La configuration technique : régression logistique (penalty='l2', C=1.0), seed=42, solver='lbfgs'.

D.1 Résultats

Détecteur	Moyenne β	Écart-type	IC 95%	β < 0
Meso4	-2.006	0.474	[-2.939, -1.089]	100%
XceptionNet	1.802	0.284	[1.243, 2.348]	0%
UCF	1.181	0.246	[0.674, 1.661]	0%
F3Net	1.190	0.259	[0.677, 1.711]	0%
Intercept	0.094	0.339	[-0.575, 0.757]	—

Tableau 11 : Statistiques bootstrap des coefficients du méta-learner (1 000 itérations)

Test statistique pour β_{Meso4} : t = -133.74, p < 0.001 (test contre $H^0 : \beta = 0$)

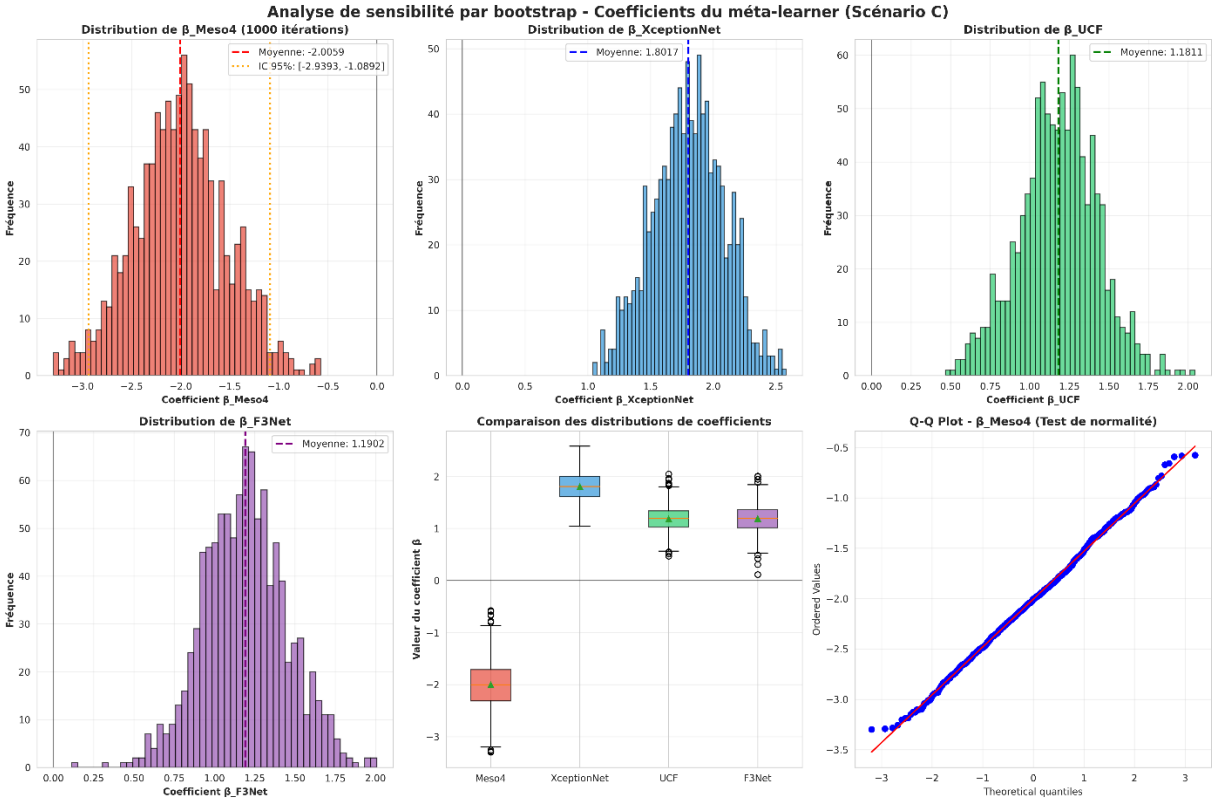

Figure 14 : Distributions bootstrap des coefficients du méta-learner

D.2 Interprétations

L'analyse bootstrap confirme trois résultats clés :

1. L'ensemble des 1 000 itérations produisent $\beta_{Meso4} < 0$, démontrant que l'inversion du signal de Meso4 est reproductible et non aléatoire.

2. L'intervalle de confiance à 95% [-2.939, -1.089] exclut complètement zéro, et le test t rejette formellement $H^0 : \beta = 0$ ($p < 0.001$).

3. Le coefficient bootstrap moyen (-2.006) diffère de seulement 0.017 du coefficient original (-1.989), ce qui valide la représentativité du Training Set malgré sa taille modeste.

Le mécanisme d'inversion de Meso4 n'est pas un artefact échantillonnal mais une propriété structurelle du méta-learner face à la distribution des erreurs de Meso4 sur les modèles de diffusion. Le méta-learner exploite activement le signal inversé de Meso4 afin de corriger le biais de calibration global, mécanisme qui sous-tend les gains de performance du Scénario C.

Annexe E : Validation statistique bootstrap des AUC

Afin de déterminer la significativité des résultats obtenus, une analyse bootstrap (1 000 itérations, rééchantillonnage stratifié, seed=42) a été conduite sur le Validation Set (n = 2 811) en travaillant exclusivement sur les scores P(FAKE) déjà calculés, sans relancer l'inférence des modèles.

E.1 Résultats

Modèle / Scénario	AUC	IC_95_low	IC_95_high	SE
Meso4	0,5017	0,4816	0,5223	0,0104
XceptionNet	0,7696	0,7523	0,7880	0,0092
UCF	0,7669	0,7509	0,7846	0,0087
F3Net	0,7513	0,7327	0,7688	0,0091
Scénario A	0,7800	0,7625	0,7972	0,0087
Scénario B	0,7826	0,7657	0,8001	0,0084
Scénario C	0,7884	0,7722	0,8052	0,0085

Tableau 12 : Intervalle de confiance 95% sur les AUC par bootstrap (1000 itérations) sur le VS

E.2 Test de significativité : Scénario C vs XceptionNet

Le test de significativité entre XceptionNet (meilleur détecteur individuel) et le scénario C (méta-learner) valide statistiquement le gain de 1.88pp sur l'AUC.

L'intervalle de confiance à 95% du delta entre les deux AUC est [0.0104, 0.0272] et présente une p-value (H^0: $\Delta AUC \leq 0$) = 0.0000. La valeur 0 est exclut de l'intervalle de confiance, ce qui démontre que dans 95% des cas la gain AUC du scénario C sur XceptionNet est situé entre 1.04pp et 2.72pp.

Annexe F : Ablation du méta-learner (analyse Meso4)

Cette annexe examine si la contribution de Meso4 au méta-learner du Scénario C provient d'une complémentarité informationnelle réelle ou d'un mécanisme de recalibration mécanique. Pour ce faire, deux variantes d'ablation ont été comparées au Scénario C original. La première consiste à évaluer un méta-learner entraîné sur trois détecteurs uniquement (sans Meso4). La deuxième remplace Meso4 par un classifieur trivial constant fixé à sa moyenne observée sur le Validation Set (P(FAKE) = 0.7126).

F.1 Résultats

Variante	AUC	Accuracy	F1	Mean_P(FAKE)	ΔAUC_vs_C
Scénario C original (4 détecteurs)	0,7884	0,6958	0,6638	0,4908	0,0000
Ablation 1 — Sans Meso4	0,7893	0,6937	0,6614	0,4912	0,0008
Ablation 2 — Meso4 constant	0,7893	0,6937	0,6614	0,4913	0,0008

Tableau 13 : Résultats des ablations du méta-learner sur le VS

Scénario	$\frac{\beta_{Meso4}}{Meso4_{const}}$	$\beta_{XceptionNet}$	β_{UCF}	β_{F3Net}	$\beta_{intercept}$
Scénario C original	-1,9887	+1,7925	+1,1870	+1,1773	+0,0881
Ablation 1 — Sans Meso4		+1,7051	+1,2018	+1,1651	-1,2912
Ablation 2 — Meso4 constant	+0,0191	+1,7017	+1,2125	+1,1548	-1,3031

Tableau 14 : Coefficients appris par le méta-learner selon les scénarios d'ablation sur le VS

F.2 Interprétations

Les trois variantes atteignent des performances quasi-identiques, l'écart maximal d'AUC étant de 0.08 pp. Ce résultat permet de tirer trois conclusions. Premièrement, la présence ou l'absence de Meso4 n'affecte pas matériellement les performances du méta-learner, ce qui exclut une complémentarité informationnelle au sens strict. Deuxièmement, le coefficient $\beta_{Meso4_{const}} \approx 0$ obtenu dans l'Ablation 2 confirme qu'une valeur constante n'apporte aucune information

discriminante exploitable. Le méta-learner compense en ajustant son intercept de +0.088 à -1.303, produisant un effet de recalibration identique par un chemin paramétrique différent.

Troisièmement, c'est exactement ce même mécanisme qui explique pourquoi l'Ablation 1 ne subit aucune chute de performance. Sans Meso4, on perd logiquement sa contribution moyenne au score linéaire (qui valait environ -1,41, soit son ($\beta_{Meso4} \times P(FAKE)_{moyen} \approx -1.99 \times 0.71$). Mais le méta-learner compense cette absence de manière totalement autonome en apprenant un intercept lourdement négatif (-1,291 contre +0,088 dans le Scénario C initial). Ce grand écart permet de restaurer un décalage global équivalent vers le seuil de 0,5.

Toutes ces observations démontrent que l'intérêt de Meso4 dans le Scénario C réside uniquement dans sa moyenne élevée (0,71). Elle agit comme un correcteur face au sous-calibrage chronique des autres détecteurs compétents, et non par la pertinence des variations de son signal. Puisque le méta-learner est tout à fait capable de reproduire cet effet de recalibration par d'autres voies paramétriques dès que Meso4 est retiré, il faut largement relativiser l'interprétation du coefficient $\beta_{Meso4} = -1.99$.

Bibliographie

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet : A Compact Facial Video Forgery Detection Network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1-7. https://doi.org/10.1109/WIFS.2018.8630761

Choi, H., & Woo, S. S. (2025). GAN or DM? In-depth Analysis and Evaluation of AI-generated Face Data for Generalizable Deepfake Detection. *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, 759-766. https://doi.org/10.1145/3672608.3707733

Chollet, F. (2017). *Xception : Deep Learning with Depthwise Separable Convolutions* (arXiv:1610.02357). arXiv. https://doi.org/10.48550/arXiv.1610.02357

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., & Verdoliva, L. (2022). *On the detection of synthetic images generated by diffusion models* (arXiv:2211.00680). arXiv. https://doi.org/10.48550/arXiv.2211.00680

Doloriel, C. T. C., Ullah, H., Liland, K. H., Machot, F. A., & Cheung, N.-M. (2026). *Towards Sustainable Universal Deepfake Detection with Frequency-Domain Masking* (arXiv:2512.08042). arXiv. https://doi.org/10.48550/arXiv.2512.08042

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale* (arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Fontana, F., Diko, A., Lanzino, R., Marini, M. R., Kaddar, B., Foresti, G. L., & Cinque, L. (2025). *Revisiting Deepfake Detection : Chronological Continual Learning and the Limits of Generalization* (arXiv:2509.07993). arXiv. https://doi.org/10.48550/arXiv.2509.07993

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks* (arXiv:1406.2661). arXiv. https://doi.org/10.48550/arXiv.1406.2661

Guarnera, L., Giudice, O., & Battiato, S. (2024). Mastering Deepfake Detection : A Cutting-edge Approach to Distinguish GAN and Diffusion-model Images. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *20*(11), 1-24. https://doi.org/10.1145/3652027

Inter, L. rédaction numérique de F. (2024, février 5). *IA : Une arnaque par deepfake a coûté 26 millions de dollars à une entreprise de Hong Kong*. France Inter. https://www.radiofrance.fr/franceinter/ia-une-arnaque-par-deepfake-a-coute-26-millions-de-dollars-a-une-entreprise-de-hong-kong-1799286

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Liu, B., Zhu, T., & Ding, M. (2025a). A Review of Deepfake and Its Detection : From Generative Adversarial Networks to Diffusion Models. *International Journal of Intelligent Systems*, *2025*(1), 9987535. https://doi.org/10.1155/int/9987535

Ma, N., Goldstein, M., Albergo, M. S., Boffi, N. M., Vanden-Eijnden, E., & Xie, S. (2024). *SiT : Exploring Flow and Diffusion-based Generative Models with Scalable Interpolant Transformers* (arXiv:2401.08740). arXiv. https://doi.org/10.48550/arXiv.2401.08740

Naskar, G., Mohiuddin, S., Malakar, S., Cuevas, E., & Sarkar, R. (2024). Deepfake detection using deep feature stacking and meta-learning. *Heliyon*, *10*(4), e25933. https://doi.org/10.1016/j.heliyon.2024.e25933

Peebles, W., & Xie, S. (2023). *Scalable Diffusion Models with Transformers* (arXiv:2212.09748). arXiv. https://doi.org/10.48550/arXiv.2212.09748

Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). *Thinking in Frequency : Face Forgery Detection by Mining Frequency-aware Clues* (arXiv:2007.09355). arXiv. https://doi.org/10.48550/arXiv.2007.09355

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake Detection : A Systematic Literature Review. *IEEE Access*, *10*, 25494-25513. https://doi.org/10.1109/ACCESS.2022.3154404

Ricker, J., Damm, S., Holz, T., & Fischer, A. (2024). *Towards the Detection of Diffusion Model Deepfakes* (arXiv:2210.14571). arXiv. https://doi.org/10.48550/arXiv.2210.14571

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models* (arXiv:2112.10752). arXiv. https://doi.org/10.48550/arXiv.2112.10752

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++ : Learning to Detect Manipulated Facial Images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1-11. https://doi.org/10.1109/ICCV.2019.00009

SCLBD. (2026). *SCLBD/DeepfakeBench* [Python]. https://github.com/SCLBD/DeepfakeBench (Édition originale 2023)

Sharma, V. K., Garg, R., & Caudron, Q. (2024). A systematic literature review on deepfake detection techniques. *Multimedia Tools and Applications*, *84*(20), 22187-22229. https://doi.org/10.1007/s11042-024-19906-1

Song, J., Meng, C., & Ermon, S. (2022). *Denoising Diffusion Implicit Models* (arXiv:2010.02502). arXiv. https://doi.org/10.48550/arXiv.2010.02502

Staněk, V., Perešíni, M., Sekanina, L., Firc, A., & Malinka, K. (2026). *Evolutionary Multi-Objective Fusion of Deepfake Speech Detectors* (arXiv:2604.01330). arXiv. https://doi.org/10.48550/arXiv.2604.01330

Stupp, C. (2019, août 30). Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. *Wall Street Journal*. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wahab, H., Ugail, H., & Jaleel, L. (2025). *Ensemble-Based Deepfake Detection using State-of-the-Art Models with Robust Cross-Dataset Generalisation* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2507.05996

Wilson, M. (s. d.). *HARKing : What is it and why is it bad?*, https://health.ucdavis.edu/media-resources/ctsc/documents/pdfs/harking-2021.pdf

Yan, Z., Zhang, Y., Fan, Y., & Wu, B. (2023). *UCF : Uncovering Common Features for Generalizable Deepfake Detection* (arXiv:2304.13949). arXiv. https://doi.org/10.48550/arXiv.2304.13949

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Chaussée de Binche 151, 7000 Mons, Belgique | **www.uclouvain.be/lsm**