COMPROMIS

COntenus Multimédia PROtégés par Machines IntelligenteS

Protected Multimedia Contents using Intelligent Machines

Projet COMPROMIS
Lot 3 : Identification

Auteurs :
Christophe Rosenberger
Patrick Bas
Teddy Furon
Nicholas Evans

Relecteur interne :
William Puech

15 avril 2026

Table des matières

1

Présentation du projet

Le projet COMPROMIS se base sur une vision moderne de la protection des données multimédia avec en son centre l'apprentissage profond. Ce projet défend l'idée que la protection des données multimédia doit nécessairement être associée à la sécurité des outils qui analysent ces données, c'est-à-dire de nos jours l'Intelligence Artificielle. Le constat est simple : La protection des données multimédia est assurément le domaine de la cybersécurité qui a le plus profité de l'I.A., mais il a négligé la vérification du niveau de sécurité de ce nouvel outil. l'IA est devenue l'un des maillons faibles de la protection des données multimédia.

Pourtant, la communauté protection des données multimédia est la plus amène pour étudier la sécurité de l'apprentissage profond en calquant ses valeurs cardinales (l'intégrité, la confidentialité et l'identification) aux données de l'apprentissage profond (données d'entraînement, données de test, modèle appris).

Cette mise en parallèle de la protection des données multimédia et de la sécurité de l'apprentissage profond crée une fertilisation croisée vertueuse assurant la sécurité de bout en bout, des données jusqu'aux algorithmes qui les traitent. Les verrous scientifiques concernent ainsi tant les applications classiques de la protection des données multimédia (analyse forensique, tatouage, biométrie, deepfakes, stéganographie, chiffrement sélectif) que le domaine émergent de l'apprentissage profond. Ces problèmes scientifiques répondent directement aux préoccupations des agences de sécurité, des industriels, et des producteurs de contenus multimédia.

Dans ce contexte et étant donné les évolutions technologiques en cours, les enjeux scientifiques auxquels ce projet entend répondre se situent dans les valeurs cardinales de la protection des données multimédia, à savoir :

1. Confidentialité : Il faut protéger la confidentialité en chiffrant les données multimédia non seulement au stockage et à la transmission mais aussi lors de leur traitement, y compris par une Intelligence Artificielle. La confidentialité s'entend aussi au sens de limitation de fuites d'information concernant des données multimédia. En effet, lorsqu'elles concernent des données personnelles (visage, voix. . .), la confidentialité protège la vie privée dans de nombreux enjeux (irréversibilité de signatures biométriques par exemple).
2. Intégrité : Cette valeur défend l'authenticité des données multimédia, à savoir s'ils ont été manipulés de manière intentionnelle en changeant le contenu sémantique ou en insérant un message caché. L'analyse de l'authenticité doit faire face à l'histoire des données, qui peuvent être éditées de multiples fois dans leur vie de manière bénigne (compression, changement de format, changement de colorimétrie...). L'enjeu est donc de discerner l'intention (détournement ou non ?) et également d'amener des preuves de manipulations.
3. Identification / Propriété : L'identification s'entend au sens large : par exemple, savoir reconnaitre une donnée multimédia (comme entant une quasi-copie d'une version originale), ou identifier sa source ou son ayant-droit (l'artiste d'une œuvre d'art). En biométrie, identifier la source revient à identifier l'individu à partir de ses traits biométriques.

2

Présentation du lot

Acteurs : LIRMM, GREYC, CRIStAL, LinkMedia, EURECOM, CEA.

L'identification s'entend au sens large : par exemple, savoir reconnaître une donnée multimédia (comme entant une quasi-copie d'une version originale), ou identifier sa source ou son ayant-droit (l'artiste d'une œuvre d'art). En biométrie, identifier la source revient à identifier l'individu à partir de ses traits biométriques. Ces enjeux scientifiques sont associés à des questions de recherche, elles-mêmes associées à des verrous scientifiques. Certains de ces verrous se justifient aussi par le nécessaire cercle vertueux de la sécurité, qui consiste à développer des attaques afin de pouvoir également concevoir des contre-attaques qui renforceront les méthodes de protection. Enfin, beaucoup de ces verrous sont motivés par des forts liens entre d'une part l'application (l'analyse forensique, la biométrie, le chiffrement sélectif,...) et, d'autre part, les nouveaux outils d'analyse, de détection, de génération fournis par l'Intelligence Artificielle. Ce lot regroupe les activités qui permettent d'identifier la personne associée à une donnée (biométrie) ou bien ses ayant droits (propriété).

2.1 Apprentissage

— T3.1 Propriété du modèle et des données d'apprentissage (resp. LinkMedia)

2.2 Applications

— T3.2 Biométrie (resp. GREYC)

1. La thèse « Génération de signatures biométriques sécurisées par des architectures profondes» (resp. GREYC) concerne la proposition d'architectures profondes génériques (indépendante de la modalité biométrique) permettant la protection de données biométriques nativement [Thèse, GREYC (C. Rosenberger), LIRMM (W. Puech)]. Elle commencera en septembre 2026 avec un stage de M2 préliminaire en mars (Titouan Le Bret).
2. La thèse « Explications de décisions en biométrie pour des applications en sécurité » (resp. GREYC) apporte des explications interprétables d'une décision en biométrie telles que pourquoi la vérification d'identité a échouée ?, ou, quelle approche a été utilisée pour réaliser une attaque par présentation détectée ? [Thèse, GREYC (C. Rosenberger, F. Jurie)]. Cette thèse a débuté en janvier 2025 (Augustin Diers),
3. La thèse « Extracteurs flous pour des données biométriques » (resp. GREYC) construit des extracteurs flous adaptés aux vecteurs de caractéristiques biométriques obtenus par des techniques d'apprentissage profond [Thèse, GREYC (P.Lacharme)]. Cette thèse a débuté en septembre 2025 (Théotime D'Oliveira),
4. Le post-doc « Campagne d'évaluation ASVspoof » (resp. EURECOM) travaille dans le cadre de la 5eme édition du concours ASVspoof sur le protocole, les métriques, la mise en œuvre et la maintenance de systèmes reproductibles. [Post-doc, Nicholas Evans, Massimiliano Todisco (EURECOM)].

— T3.3 Tatouage de contenus ou de modèles (resp. CRIStAL)

1. La thèse «Tatouage de réseau de neurones et données d'entraînement radioactives » tatoue conjointement bases d'entrainement et les paramètres du réseau.

3

Synthèse

L'apprentissage profond est une méthode essentielle dans le traitement de données multimédia. A titre d'illustration, la grande majorité des systèmes biométriques (pour identifier ou authentifier un individu) utilise un modèle profond pour la génération de paramètres biométriques pour la comparaison. L'IA peut être d'une grande aide pour concevoir des contre-mesures mais aussi pour générer des attaques. La recherche au sein du PEPR COMPROMIS dans le lot 3 se concentre sur la proposition d'attaques pour mieux concevoir en avance de phase la défense contre les attaques.

L'identification dans le domaine de l'analyse de données multimédia concerne 3 aspects.

1. **Apprentissage** : Cette brique essentielle nécessite un grand volume de données et optimise des paramètres de grands modèles d'IA. Ces modèles appris sont en général mis à disposition de la communauté scientifique pour un usage libre en recherche la plupart du temps. Malheureusement, il est connu que ces modèles sont réutilisés par des industriels posant à la fois des problèmes de propriété intellectuelle mais aussi de sécurité et respect de la vie privée. En effet, il n'est pas facile de garantir l'absence de portes dérobées (modèle répondant à une perturbation maîtrisée de l'attaquant pour modifier son comportement) ou de la possibilité de remonter aux données d'apprentissage.
2. **Biométrie** : Un des enjeux primordial en cybersécurité est la vérification de l'identité de l'individu souhaitant accéder à un service numérique (paiement, contrôle d'accès logique...). La biométrie s'est imposée comme un facteur d'authentification à la fois simple d'usage et étroitement liée à l'individu. Une donnée biométrique est une donnée multimédia (audio, image, vidéo, 3D). La biométrie n'est pas exempte d'enjeux à résoudre dont la non révocabilité intrinsèque d'une donnée biométrique, la possibilité d'attaque (comme tout système) par présentation de données altérées/falsifiées ou le possible rejeu d'une information capturée. Plusieurs contributions ont été réalisées pour répondre à ces enjeux par la proposition de protocoles cryptographiques exploitant la biométrie évitant le rejeu ou différents résultats pour la certification de systèmes biométriques. Il est en effet possible d'impacter la sécurité et la protection de la vie privée de la biométrie en proposant des propriétés souhaitées dans l'étape de certification. Des contributions sur le test de la robustesse des systèmes biométriques face à des attaques générées par IA ou la mesure de biais dans les systèmes ayant un impact sociétal mais aussi en sécurité ont été réalisées.
3. **Propriété de contenus multimédia** : A l'ère de l'IA générative (les premiers papiers sur le sujet datent de 2015), il est très difficile de savoir si un contenu multimédia est authentique ou générée. Le lot 2 du PEPR COMPROMIS se focalise notamment sur la détection de contenus générés. Une autre approche complémentaire, traitée dans le lot 3, vise à tatouer des contenus capturés (audio, photo, vidéo...) pour pouvoir prouver son authenticité (ou le fait de posséder une licence à un logiciel).

4

Présentation des principaux résultats

4.1 T3.1 Propriété du modèle et des données d'apprentissage (resp. LinkMedia)

Les grands modèles linguistiques (LLM) sont souvent affinés à l'aide d'instructions afin de les aligner sur des comportements attendus et d'améliorer leurs performances et leur capacité de généralisation. L'affinage nécessite des connaissances spécialisées pour trouver un équilibre entre diversité et qualité dans l'ensemble de données d'instructions ainsi qu'une collecte coûteuse d'annotations manuelles, en particulier pour l'alignement. Pour pallier le coût et les difficultés liés à l'affinage, les praticiens entraînent souvent leurs modèles sur des données synthétiques générées par un modèle ayant déjà reçu des instructions, tel que Bard, ChatGPT ou Claude. Cela peut également être involontaire lorsque, par exemple, des "turcs mécaniques" utilisent ChatGPT pour accomplir leurs tâches. Une telle imitation soulève des questions quant à savoir si le modèle affiné constitue une œuvre dérivée du modèle original. Dans ce contexte, il est essentiel de comprendre comment détecter les cas où les résultats des LLM sont utilisés comme données d'entraînement.

Par ailleurs, les récentes réglementations en matière d'IA imposent la transparence des modèles génératifs. Cela revêt une importance croissante dans les cas où le contenu généré pourrait être utilisé à des fins malveillantes. Une approche consiste à recourir au tatouage numérique. Elle consiste à intégrer une trace secrète dans le contenu synthétique qui peut être détectée pour identifier le modèle générateur. Dans le contexte des LLM, des techniques récentes permettent une détection efficace avec une dégradation minimale de la qualité du texte généré en modifiant l'échantillonnage des tokens suivants. Sur la base de ces deux observations, cette étude aborde la question suivante : *Que se passe-t-il lorsque l'affinage d'un LLM utilisent des données d'entraînement générés par un LLM protégé par un tatouage ?*

Nous explorons la « radioactivité » potentielle – un terme inventé par Sablayrolles et al. [2020] – du tatouage numérique des grands modèles de langage (LLM), qui désigne la capacité des données d'entraînement tatouées à contaminer un modèle. Nous examinons un modèle qui a été affiné sur un corpus susceptible de contenir du texte tatoué (voir fig. 1). La méthode de référence pour détecter la radioactivité consiste à appliquer la détection de filigrane d'origine aux sorties générées par ce modèle. Cependant, cette approche s'avère inefficace car le résidu du tatouage est un signal faible, difficilement détectable dans le texte brut en sortie. Dans ce travail, nous sommes en mesure de démontrer que le tatouage LLM est bel et bien radioactif grâce à notre protocole spécifique conçu pour révéler les traces de contamination infimes. Nos contributions sont les suivantes :

- Nous concevons des méthodes de détection de la radioactivité pour quatre scénarios basés sur l'accès au modèle (ouvert / fermé) et aux données d'entraînement (supervisées / non supervisées). Notamment, notre détection en modèle ouvert améliore les performances de plusieurs ordres de grandeur.
- Nous montrons comment obtenir des p-valeurs fiables pour la détection de tatouage sur des millions de tokens.
- Nous prouvons que le texte tatoué est radioactif dans un contexte réel où un LLM est affiné sur des données d'instructions légèrement tatouées. Par exemple, dans le scénario de modèle ouvert, nos tests détectent la radioactivité avec une p-valeur de 10^{-5} lorsque seulement 5% des données d'affinage sont tatouées.

Plus d'information ici [13].

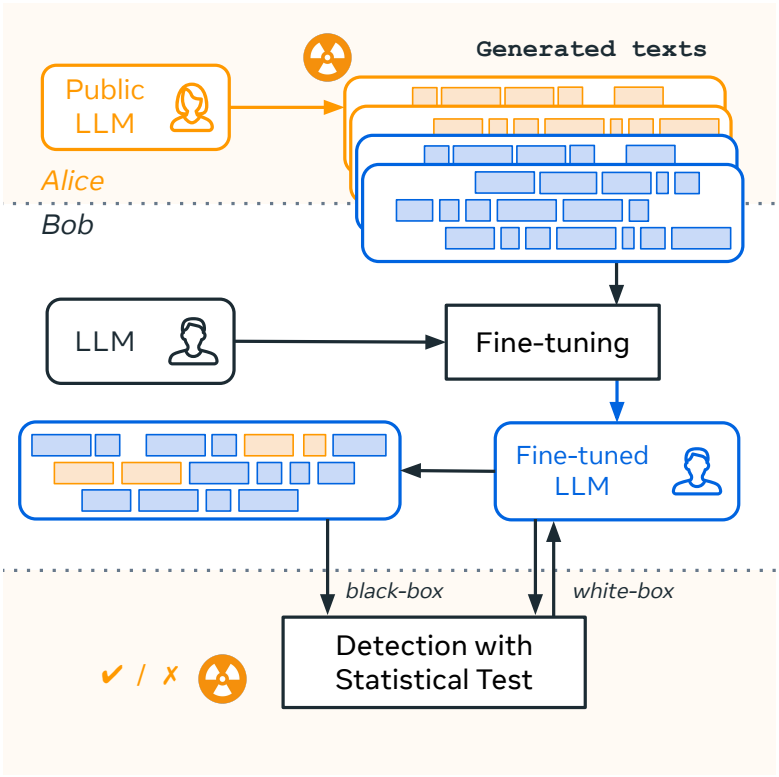


Figure 4.1 – La radioactivité du tatouage prouve que Bob a affiné son LLM sur des données d'entraînement générées par le LLM d'Alice ?

4.2 T3.2 Biométrie (resp. GREYC)

4.2.1 Protocole d'authentification sécurisée à base de biométrie

L'authentification des utilisateurs est un enjeu important sur Internet et est généralement résolue à l'aide de mots de passe statiques et souvent uniques. Une autre méthode consiste à utiliser la biométrie, mais les données biométriques sont sensibles et doivent être protégées. Des systèmes de protection tels que la génération de modèles biométriques révocables ont fait leur apparition, mais ils sont sensibles aux attaques par rejeu.

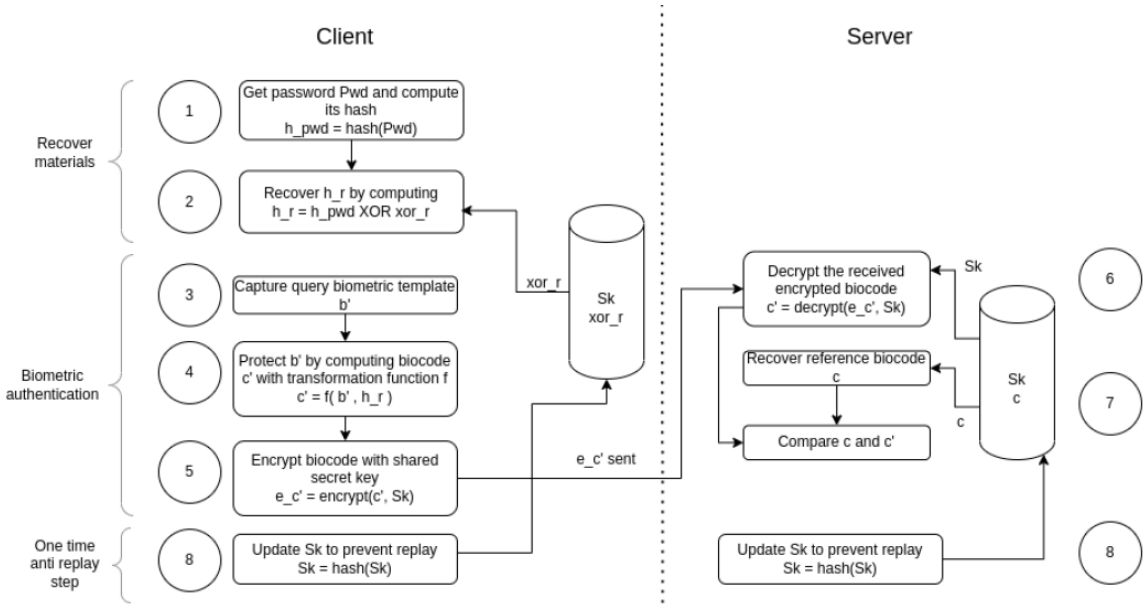


Figure 4.2 – Vérification d'identité à exploitant une donnée biométrique protégée non rejouable.

Dans cet article, nous proposons une méthode originale pour générer des modèles biométriques à usage unique pour

les applications d'authentification des utilisateurs. Le système proposé limite les attaques par rejeu, qui consistent pour un attaquant à retransmettre de manière malveillante une preuve d'identité interceptée d'un utilisateur. Notre méthode est générique : n'importe quelle modalité biométrique peut être utilisée, la vérification d'identité étant réalisée par le fournisseur de services/d'identité pour être réaliste. Les caractéristiques biométriques sont extraites des captures à l'aide de l'apprentissage profond, puis protégées par le biohashing, un système biométrique révocable. Enfin, une étape consistant en un hachage cryptographique et un chiffrement symétrique garantit la génération d'un modèle à usage unique et non reproductible. Nous avons testé notre système sur deux bases de données biométriques courantes, à partir de visages et d'empreintes digitales, et les résultats confirment son efficacité et sa robustesse face aux attaques dans le cadre d'un modèle de sécurité rigoureux [7].

4.2.2 Analyse de la sécurité des systèmes biométriques

Selon le règlement général sur la protection des données de l'UE, les données biométriques révocables sont essentielles pour protéger les modèles biométriques en combinant trois critères importants : l'irréversibilité, la révocabilité et l'impossibilité d'association. Malheureusement, de nombreux travaux ont démontré que la propriété de préservation de la distance, inhérente aux transformations, a permis de lancer des attaques basées sur la similarité. Les attaques basées sur la similarité exploitent la fuite d'informations entre la distance d'origine et la distance transformée et visent à reconstruire une caractéristique biométrique proche, utilisée pour obtenir un accès non authorisé au système.

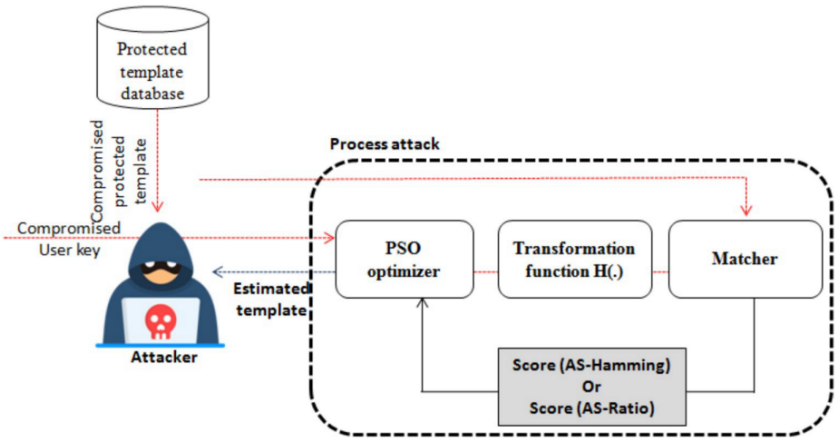


Figure 4.3 – Analyse de la sécurité d'un système biométrique à partir de la connaissance de l'attaquant (attaque par similarité).

Dans cet article, nous proposons d'atténuer les attaques par similarité en maîtrisant les connaissances de l'attaquant qui peuvent mener à leur succès. Par souci de généralité, nous reformulons les attaques par similarité pour les modèles d'ensembles non ordonnés et proposons une stratégie généralisée d'optimisation par essaim de particules pour lancer l'attaque. Nous avons souligné que le point faible permettant à l'attaque par similarité de fonctionner est le score de distance fourni par le module de correspondance. Afin de limiter les connaissances de l'attaquant, nous proposons une nouvelle stratégie de correspondance adaptée à tous les formats de modèles, basée sur le score de similarité. Nous avons réalisé des expériences et différentes comparaisons sur deux bases de données courantes, à partir d'empreintes digitales et de visages, et avons prouvé à chaque fois l'efficacité de la contre-mesure donnée face à la menace que représente l'attaque par similarité. En outre, la sécurité est abordée lorsque les connaissances de l'attaquant sont enrichies par des informations supplémentaires telles que des caractéristiques biométriques synthétiques, destinées à se rapprocher de l'espace de recherche initial. Des recommandations sont ensuite formulées afin d'atténuer ces risques au niveau de la conception [4].

4.2.3 Explication de décisions d'un système biométrique

Les systèmes biométriques sont utilisés dans notre vie quotidienne, mais ils sont susceptibles d'être contournés en tant que solution de sécurité. Les attaques par présentation dans le domaine des empreintes digitales surviennent lorsqu'un imposteur tente d'utiliser un échantillon falsifié lors de l'étape d'acquisition afin d'usurper l'identité d'une autre personne ou d'échapper à l'identification. Fournir une explication à l'opérateur (qui n'est pas un expert en biométrie) pourrait présenter un grand intérêt pour de nombreuses applications (contrôle aux frontières, contrôle d'accès physique).

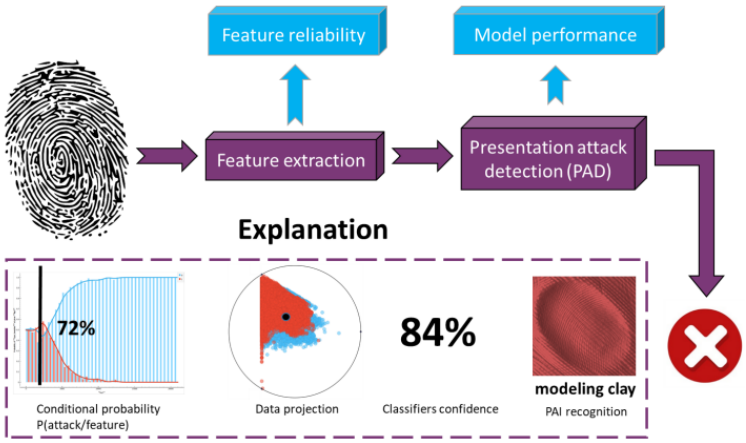


Figure 4.4 – Explication du résultat de la détection d'attaque par présentation d'empreintes digitales.

Dans cet article, nous proposons une méthode de détection des attaques par présentation d'empreintes digitales avec un retour d'explicabilité compréhensible par tout utilisateur. Les expériences ont été réalisées sur l'ensemble de données du concours Fingerprint Liveness Detection Competition (LivDet) en 2015 et contiennent plus de 58 000 images d'empreintes digitales authentiques et d'attaques. La méthode proposée atteint un taux de précision de 95,7% sur LivDet2015 avec un retour d'information compréhensible par tout utilisateur [6].

4.2.4 Évaluation d'un système de détection d'attaque par présentation

De nos jours, la biométrie est de plus en plus présente dans notre vie quotidienne. Elle est utilisée dans les documents d'identité, les contrôles aux frontières, l'authentification, les paiements électroniques, etc. Par conséquent, la sécurité des systèmes biométriques est devenue une préoccupation majeure. Le processus de certification vise à qualifier le comportement d'un système biométrique et à vérifier sa conformité aux spécifications internationales. Il implique l'évaluation des performances du système et de sa robustesse face aux attaques. Les tests de sécurité nécessitent la création d'instruments d'attaque par présentation physique (PAI), qui sont utilisés pour évaluer la robustesse des systèmes biométriques contre l'usurpation d'identité à travers de multiples tentatives de test sur l'appareil.

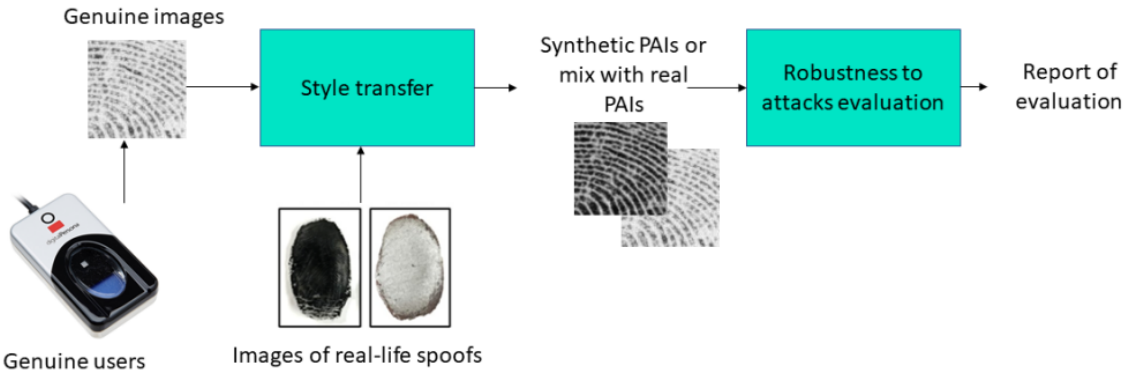


Figure 4.5 – Utilisation de données générées par IA par transfert de style pour le test de la robustesse d'un système biométrique face à des attaques par présentation.

Dans cet article, nous proposons une nouvelle solution basée sur l'apprentissage profond pour générer des images synthétiques d'empreintes digitales d'attaque à partir d'un petit ensemble de données d'images réelles acquises par un capteur spécifique. Nous modifions artificiellement ces images afin de simuler leur apparence si elles étaient générées à partir de matériaux d'attaque connus généralement utilisés dans les tests d'usurpation d'empreintes digitales. Les expériences menées sur les bases de données LivDet montrent, d'une part, que les images d'usurpation d'empreintes digitales synthétiques offrent des performances similaires à celles des images réelles du point de vue de la correspondance et, d'autre part, que les attaques par injection réussissent dans 50% des cas pour la plupart des matériaux que nous avons testés [18].

4.2.5 Estimation des biais dans les systèmes biométriques

De nos jours, la biométrie est de plus en plus présente dans notre vie quotidienne. Elle est utilisée dans les documents d'identité, les contrôles aux frontières, l'authentification, les paiements électroniques, etc. La reconnaissance faciale est de plus en plus déployée dans un large éventail d'applications, des systèmes de sécurité aux appareils personnels. Cependant, ces systèmes présentent souvent des disparités de performance importantes liées à des facteurs démographiques tels que l'âge, le sexe ou l'origine ethnique. Ces biais soulèvent de sérieuses préoccupations technologiques, éthiques et sociétales.

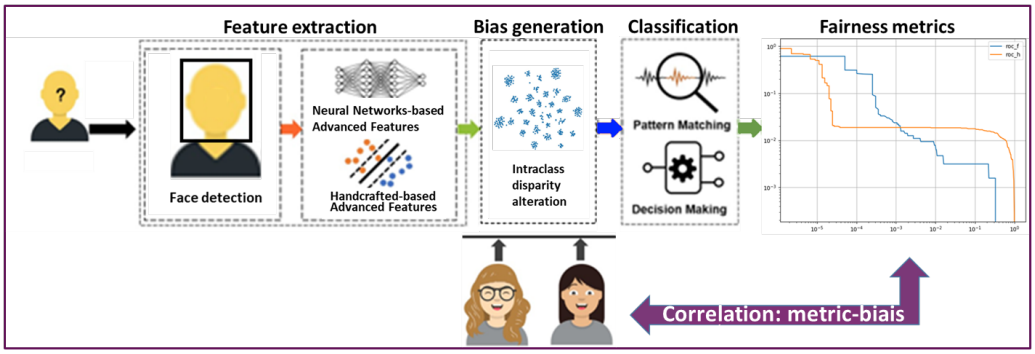


Figure 4.6 – Validation d'une métrique d'estimation de biais par analyse de la perturbation d'un groupe démographique dans l'espace des paramètres.

L'article proposé présente tout d'abord un nouveau protocole d'évaluation qui injecte systématiquement des biais algorithmiques contrôlés dans les caractéristiques biométriques extraites. Cette configuration permet de disposer d'un cadre cohérent et reproductible pour comparer les mesures d'équité existantes grâce à une analyse de corrélation, tout en garantissant que les mêmes groupes démographiques sont pris en compte dans toutes les évaluations. Nos résultats montrent que les mesures d'équité les plus largement utilisées, principalement en raison de leur formulation mathématique, ont tendance à mettre l'accent sur les disparités globales de performance entre les groupes démographiques (par exemple, asiatiques, africains, noirs), capturant ainsi les différences entre les groupes. Cependant, cela se fait souvent au détriment de la variabilité intra-groupe, c'est-à-dire les incohérences de performance entre les individus d'un même groupe démographique. Pour pallier cette limitation, nous proposons un nouvel indicateur d'équité, le TM-DP, fondé sur l'indice d'inégalité de Theil. Cet indicateur est spécialement conçu pour tenir compte des disparités de performance tant entre les groupes qu'au sein des groupes, offrant ainsi une évaluation plus nuancée et plus complète de l'équité applicable à toutes les modalités biométriques. Nous validons le TM-DP à l'aide de trois ensembles de données biométriques faciales accessibles au public, évalués à l'aide de trois extracteurs de caractéristiques différents et selon quatre seuils de décision. Les résultats expérimentaux confirment les lacunes des mesures d'équité existantes et démontrent l'efficacité et la robustesse du TM-DP [15].

4.2.6 La base de données ASVspoof 5

ASVspoof 5 est la cinquième édition d'une série de défis visant à promouvoir l'étude des attaques d'usurpation d'identité vocale et de deepfake, ainsi que la conception de solutions de détection. Nous présentons la base de données ASVspoof 5 [17], issue d'une collecte auprès de volontaires, réalisée dans des conditions acoustiques variées et auprès d'environ 2 000 locuteurs. La base contient des attaques générées à l'aide de 32 algorithmes différents, également issues d'une collecte auprès de volontaires, et optimisées à des degrés variables au moyen de nouveaux modèles de détection servant à approximer les systèmes ciblés. Ces attaques incluent des méthodes fondées sur un mélange de modèles de synthèse vocale par texte et de conversion de voix, à la fois anciens et récents, ainsi que, pour la première fois, des attaques adversariales. Les protocoles ASVspoof 5 comprennent sept partitions disjointes en termes de locuteurs : deux partitions distinctes pour l'entraînement de différents ensembles de modèles d'attaque, deux autres pour le développement et l'évaluation de modèles de détection approximatifs, et trois partitions supplémentaires correspondant aux ensembles d'entraînement, de développement et d'évaluation d'ASVspoof 5. Un ensemble auxiliaire de données collectées auprès de 30k locuteurs supplémentaires peut également être utilisé pour entraîner des encodeurs de locuteurs destinés à l'implémentation des algorithmes d'attaque. L'article décrit également une validation expérimentale de la nouvelle base de données ASVspoof 5 à l'aide de systèmes de référence pour la vérification automatique du locuteur et la détection d'attaques d'usurpation et de deepfakes. À l'exception des protocoles et outils dédiés à la génération de parole usurpée ou deepfake, l'ensemble des ressources décrites dans cet article, déjà utilisées par les participants au défi ASVspoof 5 en 2024, est désormais librement accessible à la communauté.

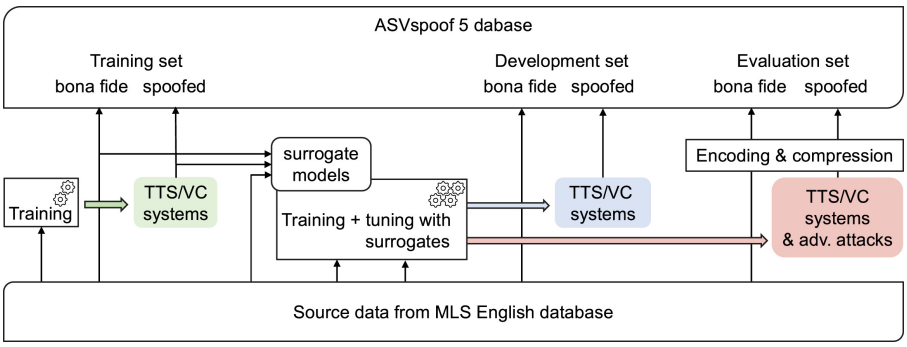


Figure 4.7 – Vue d'ensemble de la base de données ASVspoof 5. Les énoncés bona fide contenus dans les ensembles d'entraînement, de développement et d'évaluation disjoints en locuteurs proviennent directement de la base de données MLS English. Les énoncés spoofés de chacun des trois ensembles sont générés à l'aide d'algorithmes d'attaque distincts de synthèse vocale et de conversion de voix, et peuvent être combinés avec des attaques adversariales supplémentaires. Figure reproduite à partir de [17].

4.2.7 Le défi ASVspoof 5

Par rapport aux éditions précédentes, la base de données ASVspoof 5 [17] est construite à partir de données issues de la participation de volontaires, collectées auprès d'un plus grand nombre de locuteurs et dans des conditions acoustiques variées. Les attaques, elles aussi issues de la participation de volontaire, sont générées et évaluées à l'aide de modèles de détection servant à approximer les systèmes ciblés, tandis que des attaques adversariales sont intégrées pour la première fois. De nouvelles métriques permettent d'évaluer à la fois la vérification automatique du locuteur robuste aux attaques d'usurpation (SASV) et des solutions de détection autonomes, c'est-à-dire des contre-mesures indépendantes de la vérification du locuteur. Nous décrivons [16] les deux volets du défi, la nouvelle base de données, les métriques d'évaluation, les systèmes de référence et la plateforme d'évaluation, et présentons un résumé des résultats. Les attaques compromettent fortement les systèmes de référence, tandis que les soumissions apportent des améliorations substantielles.

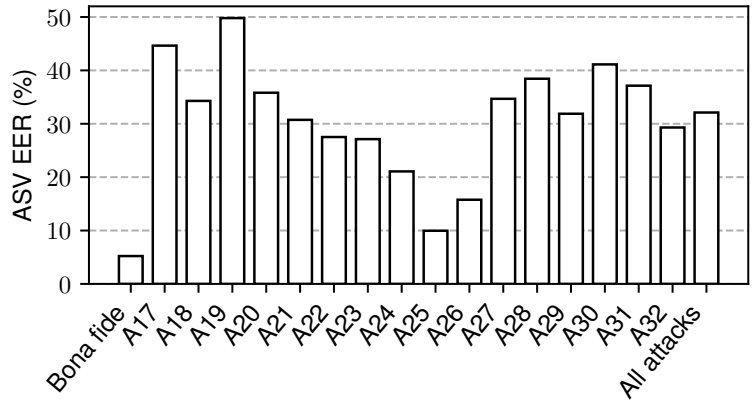


Figure 4.8 – Taux d'erreur – equal error rate (EER) – d'un système de vérification du locuteur (ASV) pour les données pour les données originales (bona fide) et les attaques dans l'ensemble d'évaluation (A17-32). Figure reproduite à partir de [16].

4.3 T3.3 Tatouage de contenus ou de modèles (resp. CRIStAL)

4.3.1 Tatouage d'images générées par un modèle de diffusion

Les modèles de diffusion sont la pierre angulaire des récentes avancées en matière de génération d'images. Des tâches autrefois complexes, telles que la génération d'images à partir de texte, la traduction d'image à image, la super-résolution ou la retouche d'images, sont désormais réalisées avec aisance. Diverses optimisations et la prolifération d'interfaces accessibles ont rendu cette technologie accessible aux utilisateurs ne disposant ni de connaissances techniques ni de matériel haut de gamme. L'IA générative crée désormais des images de haute qua-

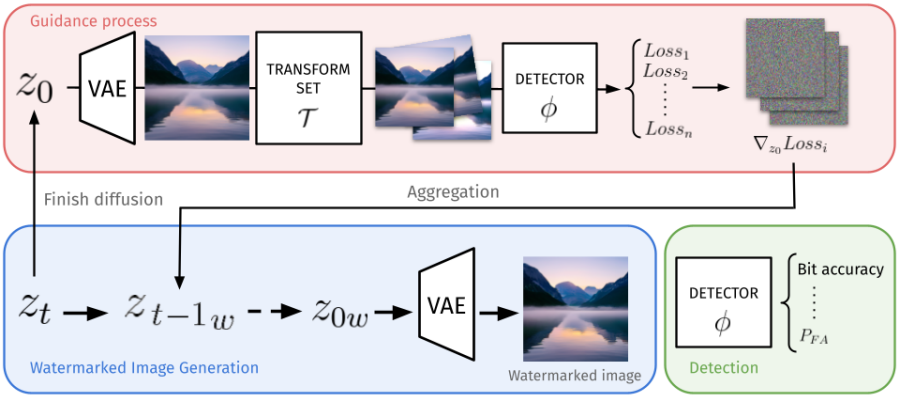


Figure 4.9 – Schéma global de la diffusion guidée par le détecteur de tatouage.

lité, variées et photoréalistes, qui sont perceptuellement impossibles à distinguer des images réelles.

Les autorités de régulation ont identifié les risques posés par cette technologie (Californian AI edit 2023 ; Chinese AI governance 2023 ; European AI Act 2023). Il existe notamment un besoin essentiel en matière d'identification et de traçabilité des contenus générés par l'IA. Parmi les solutions existantes (telles que les métadonnées (C2PA, 2024) et l'étude forensique), le tatouage numérique s'impose comme une technique clé.

Le tatouage numérique consiste à intégrer des identifiants imperceptibles dans les images, rendant celles-ci détectables par des décodeurs privés. Cette technologie éprouvée trouve de nombreuses applications, notamment la protection contre la copie, la mesure d'audience, l'identification et la monétisation de contenus. Elle a récemment été adaptée à l'identification de contenus générés par l'IA. Parmi les nombreux scénarios énumérés par la NSA, l'un consiste à avertir les utilisateurs des réseaux sociaux ou des moteurs de recherche sur Internet que ces images ne sont pas réelles, un autre consiste à filtrer les images générées par l'IA des ensembles d'apprentissage des futures IA génératives afin d'éviter un effondrement du modèle. Dans les deux cas, le détecteur de tatouage analyse des milliards d'images. La condition primordiale est un taux de fausses alertes (la probabilité de signaler une image réelle comme étant générée par l'IA) faible et vérifiable.

De nombreuses approches ont été proposées pour le texte, la voix et les images générées. Pour ce dernier type de média, la stratégie va du tatouage numérique post-génération à des modifications astucieuses du processus de génération permettant d'obtenir un contenu « intrinsèquement » tatoué. La première méthode est appelée tatouage a posteriori et la seconde tatouage intégré à la génération. Nous présentons une méthodologie permettant de convertir tout tatouage a posteriori en un schéma intégré à la génération pour tout modèle de diffusion. L'idée est d'orienter le processus de diffusion vers la génération d'images qui sont intrinsèquement considérées comme tatouées par n'importe quel détecteur de tatouage arbitraire. Nos contributions sont les suivantes :

- Notre méthode est la première à intégrer le tatouage au cours du processus de diffusion lui-même grâce à l'utilisation d'un guidage
- Elle ne nécessite aucun ré-entraînement du modèle de diffusion.
- Elle hérite de la robustesse du détecteur de tatouage, mais peut également l'améliorer face à de nouvelles attaques ciblées sans ré-entraînement du détecteur.
- Elle établit un équilibre entre la modification complète du contenu sémantique (schémas basés sur des graines) et l'ajout d'un signal invisible (schémas basés sur les VAE et schémas post-hoc).

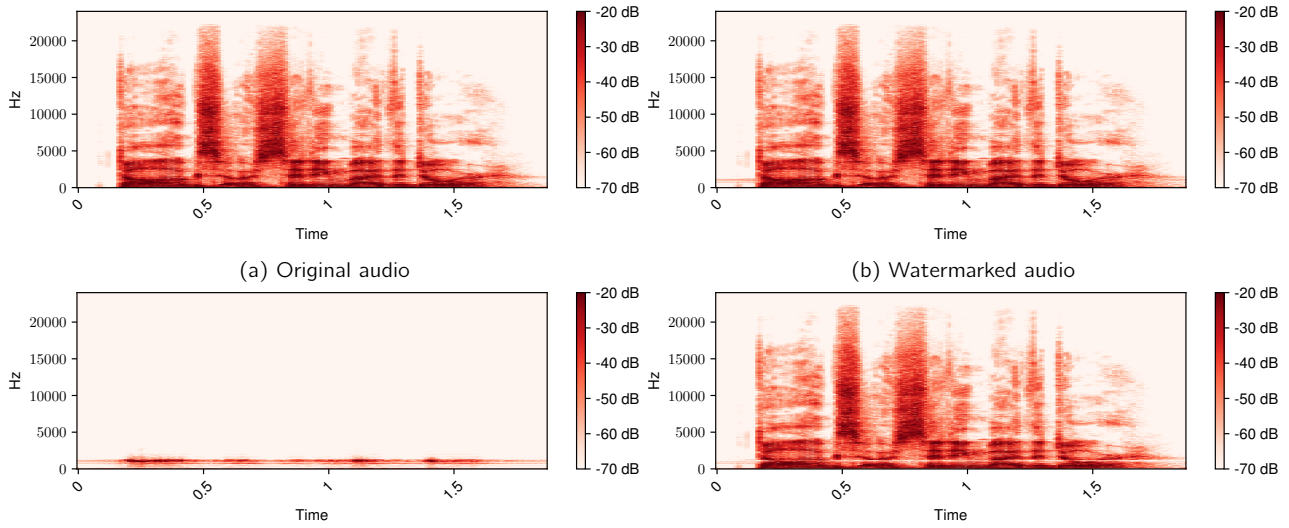
Plus d'information ici [8].

4.3.2 Analyse de sécurité d'une méthode de tatouage audio neuronal

Si le tatouage numérique existe depuis plus de 30 ans, les solutions les plus populaires actuellement reposent sur des systèmes d'apprentissages de bout en bout. En prenant comme exemple un système de tatouage à l'état de l'art permettant de localiser précisemment la génération de paroles [12], nous montrons que l'opacité du système appris permet de facilement enlever le signal de tatouage sans aucunement dégrader le contenu. Cette attaque illustre à quel point il est actuellement difficile de formaliser la contrainte de la sécurité multimédia comme une fonction de coût à optimiser. Il ne faut donc pas oublier que dans le domaine de la sécurité, les sommets de l'IA côtoient souvent des chutes d'eau. Le code source est disponible ici : https://github.com/janbutora/watermark-

anything-attack.

Plus d'informations ici [3].



(a) Original audio

(b) Watermarked audio

(c) Watermark signal

(d) Attacked audio (watermark removal)

Figure 4.10 – Analyse de sécurité d'une méthode de tatouage audio neuronal - Transformées de Fourier à court terme (STFT) des signaux originaux (a), marqués (b) et attaqués (d), ainsi que le marqueur (c). Le taux d'échantillonnage est de 48 kHz, la fenêtre de taille 1024, et pour des raisons de visualisation, le spectre de puissance maximum est limité entre -20 dB et -70 dB avec une puissance de référence de 10. Le graphique (d) montre l'effet du filtre coupe-bande pour supprimer le marqueur (zoom-in).

4.3.3 Analyse de sécurité d'une méthode de tatouage image neuronal

Le marquage par réseaux de neurones a connu un essor comme outil simple pour marquer le contenu multimédia généré de manière robuste. Cependant, cette simplicité a un coût. Bien qu'il soit facile d'imposer la robustesse dans une fonction de perte d'entraînement par diverses augmentations de données, il est actuellement inconnu comment intégrer l'aspect sécurité du marquage. Par conséquent, ces schémas en boîte noire sont facilement cassables par des attaques ciblées. Dans ce travail, nous montrons comment supprimer un marquage d'un modèle de marquage d'images WAM récemment proposé [14].

Comme le suggère le principe de Kerckhoffs, nous démontrons que l'absence de clé secrète ainsi que la robustesse requise créent un trou de sécurité facilement exploitable, représenté par des motifs quasi-périodiques qui peuvent être estimés avec précision et effacés en traitant un composant spécifique de l'image dans le domaine de Fourier. Nous montrons également que le même marquage peut être injecté manuellement, ce qui conduit à une qualité d'image bien supérieure à l'utilisation de WAM pour le marquage.

Plus d'informations ici [11].

4.3.4 Attaque par oracle

Les attaques adversaires perturbent une entrée avec une distorsion minimale afin de tromper un classifieur. La littérature examine deux contextes qui donnent lieu à deux stratégies d'attaque. Dans le contexte «boîte blanche», l'analyse de la menace part du principe que l'attaquant connaît le fonctionnement interne du classificateur. Au contraire, le cadre «boîte noire» est plus difficile sans cette connaissance.

Ce type d'attaque, connu sous le nom d'«attaques par oracle», remonte à la fin des années 1990 dans la littérature sur le tatouage numérique. En effet, un détecteur de tatouage à zéro bit n'est rien d'autre qu'un classifieur binaire. Malheureusement, la littérature récente sur l'apprentissage automatique antagoniste ne fait pas référence à ces travaux pionniers. Elle réinvente la roue, en recourant aux mêmes outils, tels que la recherche binaire, les points sensibles, l'estimation du gradient et l'approximation locale de la frontière par un hyperplan tangent. Cependant, la littérature sur l'apprentissage automatique adversaire propose également des améliorations : l'état de l'art en

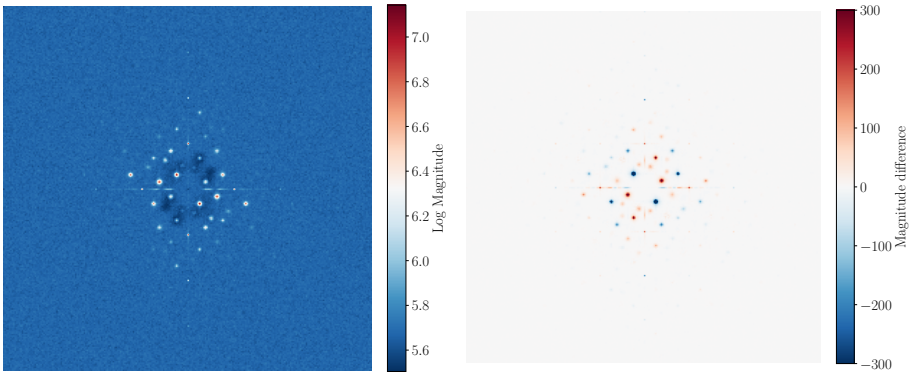


Figure 4.11 – Analyse de sécurité d'une méthode de tatouage image neuronal - Haut : Amplitude moyenne $W(m)$ du spectre de l'image marquée dans la composante de couleur du marquage, identifiée par PCA. La moyenne est calculée à partir de 500 images de couleur uniforme marquées aléatoirement. Bas : Différence entre deux amplitudes moyennes $W(m_1)$ et $W(m_2)$. Les valeurs sont limitées pour améliorer la visualisation.

matière d'attaques basées sur la prise de décision en boîte noire est le récent algorithme CGBA.

Une contre-attaque pratique est la mise à disposition non pas du vrai détecteur de tatouage, mais d'un "proxy", *i.e.* un détecteur public appris par distillation. La question est alors si une attaque montée contre le détecteur public transfère vers le détecteur privé.

4.3.4.1 Étude théorique

Un classifieur binaire induit une partition de l'espace d'entrée en régions, chacune associée à une classe donnée. Explorer l'espace entier pour dresser une carte de cette partition n'est pas faisable. L'attaquant doit interroger à maintes reprises le classifieur en boîte noire pour trouver un exemple adversaire puis réduire sa distorsion.

Les attaques de type « boîte noire » suivent un processus itératif visant à affiner la qualité de l'exemple adversaire. Le principal défi consiste à fournir des garanties théoriques, telles que la convergence vers l' exemple adversaire le plus proche. Cela se fait généralement sous certaines hypothèses fortes, comme le fait que la frontière entre les régions de classe dans l'espace d'entrée soit un hyperplan. La distorsion des exemples adversaires les plus proches trouvés dans un nombre donné de requêtes permet de contrôler le taux de convergence. Une autre question concerne le nombre de requêtes consacrées à l'estimation du gradient à chaque itération. L'estimation du gradient conduit-elle toujours à une amélioration du taux de convergence ? Cet article répond à cette question en fournissant une expression du nombre optimal de requêtes pour estimer le gradient. Cela est réalisé grâce à une justification théorique de l'approche de CGBA et à la dérivation de nouveaux résultats concernant ses performances. En particulier, nous dérivons une expression de la distorsion introduite par l'attaque à chaque itération de la recherche adversaire.

Plus d'information ici [8].

4.3.4.2 Etude pratique

L'omniprésence des contenus générés a entraîné un besoin croissant de traçabilité des contenus multimédias. Il a été démontré que les techniques de tatouage numérique offrent à la fois des garanties de détection et une grande robustesse. Cependant, une utilisation généralisée de ces méthodes nécessiterait de rendre public le détecteur de tatouage. Un tel accès compromet la sécurité du tatouage : les utilisateurs finaux disposant d'un accès illimité au détecteur pourraient facilement créer des exemples adversaires, par le biais d'attaques de type « boîte blanche » et « boîte noire ».

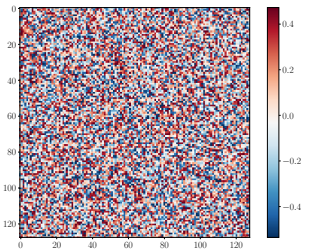
Pour contourner ce problème, nous proposons de mettre à la disposition du public un détecteur de substitution, moins précis. Les appels au détecteur privé seraient réservés aux cas importants ou anormaux. Cet article étudie les fuites d'informations potentielles provenant du détecteur de substitution. Nous créons d'abord un large panel d'images adversaires pour le détecteur de substitution. L'efficacité du détecteur privé est ensuite évaluée sur ces données. Cela nous permet d'introduire une mesure de la transférabilité de ces attaques du détecteur de substitution vers le détecteur privé. Grâce à cette mesure, nous évaluons la sécurité de différentes conceptions de détecteurs de substitution.

Plus d'information ici [9]

4.3.5 Détection du tatouage inséré par Adobe LightRoom

Dans cet article, nous abordons le problème de la détection du motif dit Adobe. Des recherches récentes [5] ont montré que les images RAW et 16 bits développées avec le logiciel Lightroom ou CameraRaw vers des formats 8 bits sont modifiées par un motif périodique imperceptible. Ce motif de taille 128×128 est influencé par le contenu en valeurs de 16 bits et est incorporé dans le domaine 16 bits, ce qui le rend impossible à estimer parfaitement à partir d'images réelles 8 bits. Comme ce motif périodique peut être perçu comme un biais partagé parmi différents utilisateurs et modèles d'appareils photo, il a conduit à des attributions de caméra inexactes lorsqu'on travaille avec la Non-Uniformité de Réponse Photographique (PRNU). Pour éliminer efficacement ce biais, il est donc nécessaire de disposer d'une méthode précise de détection du motif Adobe. Nous modélisons le motif Adobe dépendant du contenu comme un motif déterministe corrompu par un bruit uniforme, ce qui nous permet de formaliser la détection du motif Adobe comme un test d'hypothèses. En utilisant le Test du Rapport de Vraisemblance, nous démontrons que pour les images sans motif Adobe, une statistique de test soigneusement conçue suit une distribution gaussienne centrée avec une variance constante. De plus, la précision de détection dépasse 90% pour un taux de faux positifs de 10^{-4} avec des images 128×128 compressées en JPEG à une qualité de 80, et s'améliore avec une meilleure qualité d'image. Enfin, nous constatons que 16% des images du jeu de données FFHQ de visages réels contiennent le motif Adobe. Nous conjecturons que ce motif, qui est inséré par Adobe depuis au moins 2014, peut être utilisé pour de l'identification d'images ou pour une analyse forensique.

Code disponible ici : https://github.com/janbutora/adobe-detector
Plus d'informations ici [5].



(a) Motif de tatouage (taille 128×128 estimé (w).

(b) Positions du motifs sur des images portrait.

Figure 4.12 – Détection du tatouage inséré par Adobe LightRoom : présentation du signal de tatouage inséré (figure (a)) et de sa position dans l'image (figure (b)).

4.3.6 Fonction de tatouage cachée et identification de capteurs

Si l'extraction des empreintes de capteurs représente aujourd'hui un outil de police scientifique important pour l'attribution des capteurs, il a été montré récemment dans [2, 1, 10] que les images provenant de plusieurs capteurs étaient plus susceptibles de générer des Faux Positifs (FP) en présentant une "fuite" commune. Dans ce travail, nous investiguons la cause possible de cette fuite et après avoir inspecté les métadonnées EXIF des sources causant les FP, nous avons découvert qu'elles étaient liées au logiciel Adobe Lightroom ou Camera Raw. La corrélation croisée entre les résidus sur les images présentant des FP révèle des pics périodiques montrant la présence d'un motif périodique. En développant nos propres images avec Adobe Lightroom, nous avons pu montrer que tous les développement d'images brutes (ou codées sur 16 bits par canal) vers des images codées sur 8 bits intègrent également un motif périodique de taille 128×128 très similaire à un signal de tatouage. Cependant, nous montrons également que le filigrane dépend à la fois du contenu et de l'architecture utilisée pour développer l'image. Le reste de cet article présente deux méthodes différentes pour supprimer ce tatouage : l'une en le retirant de la composante de bruit de l'image, et l'autre en le retirant dans le domaine pixel. Nous montrons que pour un appareil photo présentant des FP dans [10], nous avons pu prévenir les Faux Positifs. Une discussion avec des représentants d'Adobe nous a informé que l'entreprise a décidé d'ajouter ce motif afin de provoquer un effet de "dithering".

Code disponible ici : https://github.com/janbutora/prnu-python
Plus d'informations ici [5].

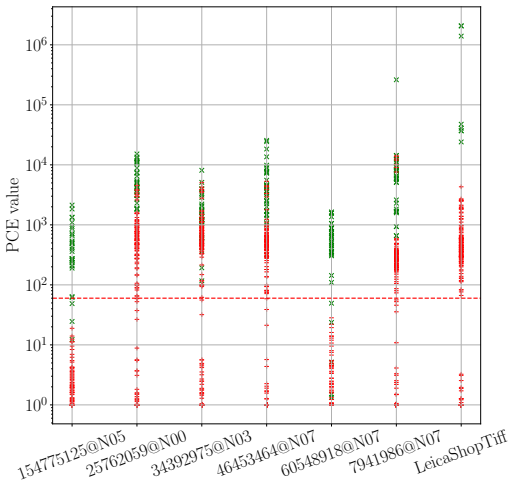


Figure 4.13 – Fonction de tatouage cachée et identification de capteurs - Les corrélations (PCE) originales pour l'appareil photo Leica Q2 et 7 appareils différents. Les six premières étiquettes correspondent aux noms d'identification FlickR, tandis que la dernière représente les images provenant d'un appareil Q2 partagé par le Leica Shop de Lille et développées avec Adobe Lightroom en format TIFF 8 bits. Les résultats des tests de correspondance sont indiqués en vert et ceux de non-correspondance en rouge. Le seuil de 60 est souligné par une ligne pointillée rouge.

Bibliographie

[1] Chiara Albisani, Massimo Iuliani, and Alessandro Piva. Checking PRNU Usability on modern devices. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2535–2539. IEEE, 2021.

[2] Daniele Baracchi, Massimo Iuliani, Andrea G Nencini, and Alessandro Piva. Facing image source attribution on iPhone X. In *Digital Forensics and Watermarking : 19th International Workshop, IWDW 2020, Melbourne, VIC, Australia, November 25–27, 2020, Revised Selected Papers 19*, pages 196–207. Springer, 2021.

[3] Patrick Bas and Jan Butora. The AI Waterfall : A Case Study in Integrating Machine Learning and Security. In *GRETSI*, Strasbourg, France, August 2025.

[4] Rima Ouidad Belguechi and Chistophe Rosenberger. Mitigate authentication attack risk on cancelable biometrics by leveraging attacker knowledge. *EURASIP Journal on Information Security*, 2025(1) :13, 2025.

[5] Jan Butora and Patrick Bas. The Adobe Hidden Feature and its Impact on Sensor Attribution. In *12th ACM Workshop on Information Hiding and Multimedia Security*, Baiona, Spain, June 2024.

[6] Augustin Diers and Christophe Rosenberger. Explainable presentation attack detection of digital fingerprints. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2024.

[7] Tanguy Gernot and Christophe Rosenberger. Robust biometric scheme against replay attacks using one-time biometric templates. *Computers & Security*, 137 :103586, 2024.

[8] Enoal Gesny, Eva Giboulot, and Teddy Furon. When does gradient estimation improve black-box adversarial attacks ? In IEEE, editor, *Proceedings of IEEE WIFS 2024*, pages 1–6, Roma, Italy, December 2024. IEEE.

[9] Chloé Imadache, Eva Giboulot, and Teddy Furon. Evaluating the security of public surrogate watermark detectors. In *Proc. of the IEEE ICASSP*, pages 1–5, Hyderabad, India, April 2025. IEEE.

[10] Massimo Iuliani, Marco Fontani, and Alessandro Piva. A leak in PRNU based source identification questioning fingerprint uniqueness. *IEEE Access*, 9 :52455–52463, 2021.

[11] Aurélien Noirault, Jan Butora, and Patrick Bas. NEURAL WATERMARKING : LACK OF A SECRET KEY IS STILL LACK OF SECURITY. working paper or preprint, September 2025.

[12] Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre Défossez, Teddy Furon, and Tuan Tran. Proactive detection of voice cloning with localized watermarking. In *Proceedings of the 41st International Conference on Machine Learning*, pages 43180–43196, 2024.

[13] Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking Makes Language Models Radioactive. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*., volume Spotlight, pages 1–35, Vancouver, Canada, December 2024.

[14] Tom Sander, Pierre Fernandez, Alain Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In *International Conference on Learning Representations (ICLR)*, 2025.

[15] Kaira Neily Sanon, Joël Di Manno, Christophe Charrier, and Christophe Rosenberger. A new fairness evaluation metric of biometric systems based on the theil inequality. In *57th International Carnahan Conference on Security Technology (ICCST 2025)*, 2025.

[16] Xin Wang, Hector Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi. ASVspoof 5 : crowdsourced speech data, deepfakes, and adversarial attacks at scale. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pages 1–8, 2024.

[17] Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, Junichi Yamagishi, Myeonghun Jeong, Ge Zhu, Yongyi Zang, You Zhang, Soumi Maiti, Florian Lux, Nicolas Müller, Wangyou Zhang, Chengzhe Sun, Shuwei Hou, Siwei Lyu, Sébastien Le Maguer, Cheng Gong, Hanjie Guo, Liping Chen, and Vishwanath Singh. Asvspoof 5 : Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *Computer Speech & Language*, 95 :101825, 2026.

[18] Abdarahmane Wone, Joël Di Manno, Christophe Charrier, and Christophe Rosenberger. Fingerprint spoof generation using style transfer. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2025.