UNIVERSITÉ DU QUÉBEC À MONTRÉAL

CONCEPTION D'ATTAQUES PAR INFÉRENCE D'APPARTENANCE CONTRE LES DONNÉES GÉNOMIQUES DANS LE MODÈLE BOÎTE NOIRE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

MOHADESEH BAGHERI

NOVEMBRE 2025

REMERCIEMENTS

Je tiens à adresser mes plus sincères remerciements à mon directeur de mémoire, Marc-Olivier Killijian, pour son accompagnement rigoureux, ses conseils avisés et son soutien constant tout au long de ce travail. Sa disponibilité et sa bienveillance ont été essentielles à la réalisation de ce projet.

Je remercie également ma famille pour leur patience, leur amour et leur soutien moral inconditionnel, qui m'ont permis d'avancer avec confiance, même dans les moments les plus difficiles.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

UQAM Université du Québec à Montréal.

AA Apprentissage Automatique.

GWAS Genome-wide association studies.

NIH National institutes of health.

HIPAA Health Insurance Portability and Accountability Act.

NHGRI National Human Genome Research Institute.

SNP Single-nucleotide polymorphism.

ADN Acide désoxyribonucléique.

ML Machine learning.

MLAAS Machine learning as a service.

MIA Membership inference attack.

HGP Human genome project.

CNN Convolutional neural network.

SVM Support Vector Machine.

XGBOOST eXtreme Gradient Boosting.

RELU Rectified linear unit.

SGD Stochastic gradient descent.

DP-SGD Differentially-private stochastic gradient descent.

PATE Private Aggregation of Teacher Ensembles.

API Application programming interface.

DNN Deep neural network.

RF Random Forest.

LR Logistic regression.

TP True positive.

TN True negative.

FP False positive.

FN False negative.

TPR True positive rate.

FPR False positive rate.

ROC Receiver operating characteristic.

AUC Area under the curve.

MLP Multilayer Perceptron.

NOTATION

Variables

x_i profil génétique (échantillon d'entrée) avec d SNPs.

y_i phénotype binaire associé à x_i (par exemple, résistance ou sensibilité).

$f(x)$ modèle cible entraîné sur (x_i, y_i).

$f_s(x)$ modèle d'ombre entraîné sur des données auxiliaires (x_i, y_i^s).

$\hat{y}$ sortie du modèle : score de confiance associé à la classe prédite.

$\mathcal{A}$ modèle d'attaque entraîné à distinguer les membres et non-membres.

$D_{\textbf{train}}$ sous-ensemble d'entraînement du modèle cible (membres).

$D_{\textbf{test}}$ sous-ensemble de test du modèle cible (non-membres).

$D_{\textbf{shadow}}$ données utilisées pour entraîner les modèles d'ombre (autres phénotypes ou jeux externes).

D'_k k^e jeu d'entraînement d'un modèle d'ombre f_s^k.

T'_k jeu de test associé à D'_k, pour générer des prédictions non-membres.

P_k^m ensemble de prédictions de f_s^k sur D'_k (membres).

P_k^n ensemble de prédictions de f_s^k sur T'_k (non-membres).

$D_{\textbf{attaque}}$ exemples membres/non-membres utilisés pour entraîner $\mathcal{A}$.

$[p_1, ..., p_n]$ vecteur de probabilités en sortie du modèle (logits normalisés).

Top-k les k plus grandes valeurs dans le vecteur $[p_1, ..., p_n]$.

RÉSUMÉ

Avec l'apparition du séquençage à haut débit et l'intégration de l'intelligence artificielle, de nouvelles préoccupations liées à la vie privée ont émergé. Les données génétiques humaines, en particulier, révèlent des prédispositions aux maladies et des éléments héréditaires familiaux. Contrairement aux données classiques, les données génomiques sont uniques, immuables et personnelles. Cette spécificité les rend particulièrement vulnérables aux abus en cas de fuite ou de mauvaise gestion. Dans ce contexte, les attaques par inférence d'appartenance (membership inference attacks – MIA) représentent une menace croissante : elles permettent à un adversaire de déterminer si un échantillon spécifique a été utilisé pour entraîner un modèle d'apprentissage automatique, compromettant ainsi la confidentialité des données biomédicales.

Ce mémoire s'inscrit dans une démarche de sensibilisation aux risques liés à la vie privée dans les applications d'apprentissage automatique sur des données génomiques. Il vise à évaluer la robustesse des modèles prédictifs lorsqu'ils sont exposés à des attaques d'inférence d'appartenance, en considérant deux méthodologies réalistes. La première repose sur la création de modèles d'ombre dans un espace de distribution similaire à celui du modèle cible, mais en s'appuyant sur des phénotypes biologiquement corrélés. Cette stratégie exploite la proximité fonctionnelle entre certains traits mesurés pour améliorer l'efficacité de l'attaque, tout en supposant un accès partiel à des données de même nature. La seconde méthodologie adopte une approche plus générique, fondée sur la généralisation des connaissances : des modèles d'ombre sont formés sur des jeux de données hétérogènes, sans similarité directe avec le modèle cible, ce qui reflète un scénario plus réaliste et contraint. La contribution principale de ce mémoire est la mise en œuvre et l'évaluation de ces deux méthodologies d'attaque MIA appliquées aux données génétiques. Afin d'évaluer la pertinence et l'efficacité de ces approches, nous avons recours à un jeu de données génomiques de levure, en raison de sa disponibilité publique et de son usage en recherche génomique. Ce jeu de données permet de simuler des expériences reproductibles et représentatives tout en contrôlant les variables biologiques pertinentes.

Les résultats expérimentaux obtenus mettent en évidence la faisabilité d'attaques par inférence d'appartenance même en l'absence totale d'informations sur les données d'entraînement du modèle cible. Les deux méthodologies proposées montrent des performances élevées, en particulier dans la détection des échantillons membres. Ces constats soulignent l'importance de développer des mécanismes de défense plus robustes et adaptés aux spécificités des données génomiques. Ils révèlent également que la sécurité des modèles d'apprentissage automatique dans le domaine biomédical ne peut être assurée uniquement par la limitation de l'accès aux données, mais qu'elle nécessite aussi des garanties algorithmiques.

INTRODUCTION

Ces dernières années, l'intersection entre l'apprentissage automatique (AA) et la génomique a profondément transformé la recherche biomédicale et la médecine personnalisée. Les technologies de séquençage à haut débit ont permis d'explorer en profondeur les relations complexes entre le génotype — c'est-à-dire la composition génétique d'un individu — et le phénotype, c'est-à-dire les caractéristiques observables telles que la taille, la susceptibilité à une maladie ou la réponse à un traitement.

Parmi les méthodes d'analyse les plus répandues dans ce domaine, les études d'association pangénomique, *Genome-wide association studies (GWAS)* [1], occupent une place centrale. Cette méthode statistique permet d'identifier les loci génétiques associés à une caractéristique donnée, en analysant la fréquence de certains polymorphismes nucléotidiques simples (*Single-nucleotide polymorphisms (SNPs)*), c'est-à-dire des variations portant sur un seul nucléotide à une position précise du génome, dans de larges cohortes d'individus (Wright et Fessele, 2017). Les résultats des GWAS servent de point de départ pour la sélection de caractéristiques et facilitent l'élaboration de modèles plus simples. En réduisant la dimensionnalité, ils améliorent à la fois l'efficacité informatique et la valeur biologique des prédictions. Cependant, la diffusion publique des résultats de ces études peut également entraîner des risques accrus pour la vie privée, en exposant indirectement des informations personnelles sur les participants.

À titre d'exemple, l'étude pionnière de Homer *et al.* (2008) a montré qu'il était possible d'identifier la présence d'une personne dans une base de données GWAS agrégée en comparant ses données génétiques aux fréquences alléliques publiées, où un allèle désigne l'une des versions d'une séquence d'ADN à un locus donné, généralement hérité de chaque parent. Cette capacité à détecter la présence d'un individu, même à partir de données statistiques globales, a soulevé des inquiétudes majeures concernant la confidentialité. Après cette révélation, les *National Institutes of Health (NIH)* — l'agence fédérale américaine de recherche biomédicale — ont restreint l'accès public à certaines

1. Les GWAS sont des études visant à identifier des associations statistiques entre des variations génétiques (comme les SNPs) et des traits phénotypiques ou des maladies, en analysant l'ensemble du génome d'un grand nombre d'individus (Visscher *et al.*, 2017)

bases de données génomiques et transféré des statistiques agrégées de GWAS sous un régime d'accès contrôlé, afin de limiter les risques de ré-identification (Zerhouni et Nabel, 2008).

La capacité des modèles à prédire des informations sensibles soulève des enjeux majeurs pour la protection des données personnelles. Les données génomiques sont irrévocables, propres à chaque individu et renferment des informations héréditaires. L'utilisation de ces données pour former des modèles d'apprentissage automatique peut entraîner des fuites d'informations, surtout si les modèles sont mis en œuvre dans des environnements accessibles au public ou aux chercheurs.

Parmi les menaces identifiées, les attaques par inférence d'appartenance (*Membership inference attack (MIA)*) ont suscité une attention croissante. Dans ce type d'attaque, l'adversaire cherche à déterminer si un échantillon a servi à l'entraînement d'un modèle. Elle exploite les variations de comportement du modèle entre les exemples vus (membres) et ceux non vus (non-membres), surtout en cas de surapprentissage. Le surapprentissage désigne la situation où un modèle apprend trop fidèlement les particularités (et le bruit) des données d'entraînement, au détriment de sa capacité de généralisation. Il se manifeste par un écart notable entre les performances d'entraînement et de test, ainsi que par des réponses surconfiantes sur les exemples vus, ce qui accroît la séparabilité membre/non-membre exploitée par les MIAs. En ce qui concerne les données génomiques, cette fonctionnalité pourrait révéler la participation d'une personne à une étude médicale ou son lien avec une information sensible, ce qui pose un risque majeur pour la confidentialité.

Dans des domaines tels que la vision par ordinateur et le traitement du langage naturel, les attaques MIA ont montré que les modèles peuvent mémoriser des données sensibles, même involontairement. Shokri *et al.* (2017) ont montré que des modèles surappris permettent d'inférer l'appartenance d'un échantillon avec une précision élevée. Les modèles de plongement, qui transforment des entités discrètes (mots, k-mers ou catégories) en un espace vectoriel dense où la proximité reflète des régularités statistiques, sont utiles mais peuvent, comme l'ont montré Song et Raghunathan (2020), mémoriser et restituer des paires mot-contexte sensibles vues à l'entraînement. Enfin, Carlini *et al.* (2021) ont montré que des modèles de langage peuvent régénérer mot pour mot des séquences confidentielles vues à l'entraînement. Ces résultats soulignent l'ampleur du risque et la nécessité d'étudier ces attaques en génomique. Si ces attaques sont bien documentées dans d'autres domaines, leur transposition à

la génomique reste marginale. Les recherches sur les MIA appliquées à la génomique sont limitées en raison du manque de données publiques combinant génotypes de qualité et phénotypes fiables, notamment du fait de contraintes éthiques, juridiques et de confidentialité (Gymrek *et al.*, 2013; Erlich et Narayanan, 2014). En particulier, des lois comme le HIPAA (*Health Insurance Portability and Accountability Act*) aux États-Unis ou le RGPD (Règlement général sur la protection des données) en Europe imposent des restrictions strictes sur l'accès aux données génétiques, car elles sont permanentes, difficilement anonymisables et potentiellement identifiables. Enfin, la forte dimensionnalité des données génomiques — où le nombre de *SNPs* dépasse largement celui des échantillons — accentue le risque de surapprentissage.

Pour contourner ces limitations, tout en conservant un cadre expérimental réaliste, nous avons choisi d'utiliser des données génomiques issues de Saccharomyces cerevisiae (la levure). Les données de levure ont été choisies pour ce projet en raison de leur diversité génétique suffisante, de leur accessibilité libre ainsi que de leur annotation précise. Ces caractéristiques en font un cadre expérimental idéal pour évaluer les attaques par inférence d'appartenance sur des données génomiques réelles, tout en évitant les contraintes éthiques et juridiques liées aux données humaines (Skelly *et al.*, 2013).

La plupart des recherches actuelles partent du principe qu'un adversaire connaît la structure interne du modèle ciblé, ce qui correspond à un scénario en boîte blanche. Dans la réalité, les adversaires n'ont souvent accès qu'aux sorties du modèle, comme dans les services d'apprentissage automatique en tant que service (*Machine learning as a service (MLAAS)*), ce qui correspond à un cadre en boîte noire. Dans ce contexte, concevoir une attaque efficace est nettement plus difficile (Truex *et al.*, 2019).

Ce mémoire explore la possibilité d'une attaque par inférence d'appartenance sur des données génomiques dans un contexte réaliste. L'objectif principal est de démontrer qu'un adversaire peut déterminer si un échantillon a été utilisé pour l'entraînement d'un modèle cible, sans connaître sa structure ni accéder à ses paramètres internes. Pour ce faire, l'étude se concentre sur la création d'un modèle d'ombre généralisable, formé sur un ensemble de données distinct, mais capable d'imiter efficacement le comportement du modèle cible. Un modèle d'ombre est un classifieur entraîné par l'adversaire pour reproduire le comportement du modèle cible. En générant, via ces modèles, un jeu de sorties annotées membre/non-membre, l'adversaire entraîne ensuite un modèle d'attaque binaire capable d'inférer

l'appartenance à partir des seules sorties du modèle cible en boîte noire (ou de ses états internes en boîte blanche). Pour évaluer la performance des attaques, plusieurs méthodologies seront testées et comparées, en mettant l'accent sur leur capacité à détecter précisément les membres (vrais positifs) sans augmenter le nombre de faux positifs. Cette analyse compare différentes architectures de modèles d'ombre pour déterminer les plus efficaces et les plus applicables. Opérationnellement, nous construisons un jeu d'évaluation contrôlé membres/non-membres et reportons des métriques standard (AUC, précision, TPR@FPR) avec intervalles de confiance sur plusieurs répétitions.

Contrairement à l'étude de Chen *et al.* (2020), qui applique des attaques MIA dans un cadre en boîte blanche à l'aide de données génomiques de levure, notre approche explore une situation plus réaliste en boîte noire, dans laquelle l'attaquant n'a accès qu'aux sorties du modèle cible. Nous proposons également une stratégie de généralisation fondée sur des modèles d'ombre entraînés à partir de phénotypes biologiquement liés ou de jeux de données hétérogènes. Même si le cadre est plus contraignant, nos résultats sont meilleurs que ceux de l'étude de Chen *et al.* (2020). Cela prouve l'efficacité et la robustesse de notre approche. La nouveauté de ce travail réside dans l'évaluation d'attaques MIA en boîte noire appliquées aux données génomiques, avec deux méthodologies complémentaires (corrélation biologique et transfert généralisé), ce qui n'a pas encore été étudié dans ce contexte.

Ce projet vise à démontrer qu'il est possible, même dans un cadre en boîte noire, d'extraire des informations sensibles sur la participation d'un individu à une étude génomique, uniquement à partir des sorties d'un modèle d'apprentissage automatique. En révélant la vulnérabilité de modèles déployés dans des contextes réalistes, ce travail souligne l'urgence de repenser les pratiques de publication, de partage et de protection des données génétiques. Ces résultats ont des implications majeures, tant pour le développement de modèles robustes et responsables que pour la confiance du public dans la recherche biomédicale et la gouvernance éthique des données. Comme l'ont montré McGuire *et al.* (2008), le séquençage du génome entier pose des défis spécifiques en matière de confidentialité, car il peut révéler des informations personnelles difficilement anonymisables.

Ce mémoire est structuré comme suit :

— Chapitre 1 – Notions préliminaires en génomique et apprentissage automatique : ce chapitre introduit le contexte général de l'étude, les enjeux liés à la confidentialité des données génomiques, les motivations scientifiques et éthiques du projet, ainsi que la problématique principale centrée sur les attaques par inférence d'appartenance dans un cadre réaliste.

— Chapitre 2 – État de l'art : il présente une revue des travaux existants sur les *MIA*s, les méthodes d'apprentissage automatique appliquées à la génomique, les approches de protection de la vie privée, ainsi que les défis spécifiques liés à la dimensionnalité élevée et au manque de jeux de données publics dans ce domaine.

— Chapitre 3 – Méthodologie : ce chapitre présente en détail les deux approches méthodologiques développées dans le cadre de ce projet. La première, appelée attaque par modèles d'ombre corrélés, repose sur l'entraînement de modèles d'ombre à partir de phénotypes auxiliaires biologiquement corrélés au phénotype cible. Cette méthode suppose que l'adversaire a accès à un sous-ensemble de données appartenant au même espace de distribution que celles du modèle cible, bien que les étiquettes soient différentes. La seconde approche, dite attaque par transfert généralisé, s'inspire des travaux de Salem et collaborateurs et consiste à former des modèles d'ombre sur des jeux de données totalement hétérogènes, sans lien direct avec le domaine génomique cible. Dans ce cas, les vecteurs de sortie des modèles d'ombre sont transformés en caractéristiques statistiques, telles que les valeurs top-k, afin d'alimenter un modèle d'attaque entraîné indépendamment.

— Chapitre 4 – Résultats et analyse : ce chapitre expose les résultats expérimentaux obtenus pour chaque méthodologie d'attaque testée, en comparant les performances des modèles selon différents critères (précision, AUC (*Area under the curve (AUC)*), taux de vrai positif, taux de faux positif). Il propose une analyse critique des résultats, identifie les limites du cadre expérimental et discute de l'impact potentiel des conclusions sur la sécurité des modèles d'apprentissage dans le domaine génomique.

— Chapitre 5 – Conclusion : ce dernier chapitre récapitule les contributions principales du travail, met en lumière les implications de ces résultats sur la confidentialité des données génomiques et propose des pistes pour des recherches futures, notamment l'adaptation de l'approche à d'autres types de données biologiques, l'extension à des ensembles de données plus volumineux ainsi que l'exploration de nouvelles architectures de modèles d'attaque ou de généralisation.

CHAPITRE 1
NOTIONS PRÉLIMINAIRES EN GÉNOMIQUE ET APPRENTISSAGE AUTOMATIQUE

La révolution génomique, combinée à l'accès croissant aux données génétiques, a permis aux modèles d'apprentissage automatique de transformer profondément le domaine de la génomique. Ces avancées ont transformé des domaines clés tels que la médecine personnalisée, la recherche pharmaceutique et l'épidémiologie. Toutefois, la sensibilité et la durabilité des données génétiques posent des défis majeurs en matière de protection de la vie privée et de sécurité. De plus, les modèles d'apprentissage automatique peuvent mémoriser certaines données d'entraînement, ce qui peut être exploité par des adversaires pour révéler des informations sensibles. Dans ce chapitre, nous présentons les concepts fondamentaux liés aux données génomiques, leur importance et les défis associés à leur protection. Nous examinons également l'utilisation de l'apprentissage automatique dans ce domaine ainsi que les menaces à la confidentialité, en particulier l'attaque par inférence d'appartenance.

Exemple récent : fuite de données chez 23andMe. Un exemple récent illustrant le rôle crucial du niveau de connaissance de l'adversaire est la fuite de données survenue chez 23andMe en 2023. Même si cette attaque ne correspondait pas à une attaque d'inférence d'appartenance, elle démontre clairement comment des informations variées peuvent servir à un attaquant pour compromettre la confidentialité de données sensibles, comme celles contenues dans le génome.

Dans ce cas, l'attaquant a lancé une attaque par bourrage d'identifiants (credential stuffing), en profitant du fait que de nombreux utilisateurs réutilisent les mêmes mots de passe sur plusieurs plateformes. Cette stratégie repose sur une compréhension préalable du comportement des utilisateurs. De plus, l'attaquant a utilisé des informations d'identification obtenues lors de violations de données précédentes, ce qui lui a permis de les tester massivement sur 23andMe, exploitant ainsi l'absence de contrôle de limitation de tentatives dans l'API de connexion du site. Après avoir compromis quelques comptes, l'attaquant a pu profiter des fonctionnalités sociales de la plateforme, telles que la recherche de correspondances ADN et les arbres généalogiques [1] partagés, pour accéder aux données intercon-

1. Un arbre généalogique est une représentation schématique des liens de parenté entre individus, permettant de retracer les relations familiales sur plusieurs générations.

nectées de plusieurs milliers d'autres utilisateurs. Cette attaque démontre une compréhension structurelle de la plateforme ciblée et une connaissance de la valeur des données. Par exemple, en ciblant des utilisateurs d'ascendance ashkénaze juive ou chinoise, ou ceux associés à des personnes fortunées.

Ce scénario met en évidence le fait qu'un adversaire bien informé peut exploiter des failles de sécurité, des comportements humains et des logiques de systèmes pour mener une attaque à grande échelle, même avec un accès limité. Cela renforce l'idée que, dans un cadre d'inférence d'appartenance, le niveau de connaissance de l'adversaire est un facteur critique pour la réussite de l'attaque, qu'il s'agisse de connaître la distribution des données, les sorties du modèle ou la structure du système (Holthouse *et al.*, 2025). Ces exemples motivent l'étude, dans les chapitres suivants, des attaques plus subtiles qui exploitent les modèles d'apprentissage automatique eux-mêmes, comme les attaques par inférence d'appartenance.

1.1 Notions préliminaires en génomique

Avec l'évolution des technologies de séquençage, les données génomiques sont devenues une ressource essentielle dans les domaines médicaux, notamment pour la médecine de précision, la recherche génétique et la modélisation prédictive. Le séquençage de l'*Acide désoxyribonucléique (ADN)* est un processus de laboratoire qui permet de cartographier la séquence complète du génome d'un individu. Ce procédé a été initié en 1990 par le *National institutes of health (NIH)*, et le premier séquençage a été obtenu après treize années en dépensant trois milliards de dollars. Cependant, au fil du temps, les coûts et les délais ont fortement diminué : aujourd'hui, des protocoles de séquençage du génome entier rapides ou ultra-rapides permettent un retour de résultats en quelques jours (*médiane* $\sim$2,3 jours pour l'ultra-rapide) (Kansal, 2025), tandis que le coût par génome a chuté de façon marquée au cours de la dernière décennie, selon les séries du *National Human Genome Research Institute (NHGRI)* (Wetterstrand, 2023). Dans ce qui suit, nous rappelons brièvement la structure et la réplication de l'ADN afin de situer la nature des données produites par ces technologies.

Sur le plan moléculaire, la compréhension de la double hélice éclaire la manière dont l'information génétique est lue et copiée. L'ADN est une molécule formée de deux brins complémentaires en double hélice, dont la séquence de quatre bases nucléotidiques – l'adénine (A), la thymine (T), la cytosine

(C) et la guanine (G) – porte l'information génétique. Ces bases nucléotidiques s'associent spécifiquement (A – T et C – G), fournissant un gabarit complémentaire pour la réplication. Toutefois, la haute fidélité ne provient pas du seul appariement : elle résulte également de la sélectivité des ADN polymérases, de l'activité d'exonucléase 3'→5' (proofreading) – y compris des mécanismes de *proofreading extrinsèque*(extrinsic proofreading) – et de la réparation des mésappariements (MMR) agissant au niveau de la fourche de réplication (Zhou et Kunkel, 2022). Ces mécanismes de fidélité conditionnent directement la qualité des lectures et l'interprétabilité des jeux de données issus du séquençage. Enfin, l'ordre et la composition des nucléotides (p. ex. le contenu GC) influencent l'expression, la régulation et la fonction ; leurs effets ne se limitent pas aux protéines, mais concernent aussi les ARN non codants et les éléments régulateurs (Grome et Isaacs, 2021). Le séquençage de l'ADN consiste à déterminer l'ordre exact de ces nucléotides. Depuis la découverte de la structure de l'ADN, diverses technologies de séquençage ont été développées pour décoder efficacement les informations génétiques et, aussi, les technologies récentes ont permis la génération massive de données de séquençage pour différentes espèces (Wong *et al.*, 2019).

FIGURE 1.1 – Illustration schématique de la séquence d'ADN composée de quatre bases (A, T, C, G). Oestreich *et al.* (2021).

Sur cette base moléculaire, nous passons aux unités fonctionnelles et aux niveaux de variation pertinents pour l'analyse génomique. Un gène est une unité fonctionnelle de l'ADN dont les produits peuvent être une protéine ou un ARN fonctionnel (p. ex. ARNt, ARNr, microARN, ARN long non codant), et dont les régions régulatrices contrôlent où et quand ces produits sont exprimés. Bien que de nombreux gènes soient communs chez l'humain, des variations existent sous forme d'allèles, hérités de chaque parent. Ces différences entre individus sont appelées polymorphismes génétiques. Le type le plus fréquent est le SNP, qui implique une différence dans un seul nucléotide à une position spécifique du génome. Les SNPs peuvent influencer l'expression des gènes et le fonctionnement des protéines, affectant ainsi les phénotypes, c'est-à-dire les caractéristiques observables. Établir la relation entre les SNPs et les phénotypes est essentiel pour identifier les facteurs génétiques associés aux

maladies (Botta *et al.*, 2014; Wright et Fessele, 2017). Les Figures 1.2 et 1.3 illustrent respectivement ces notions de locus/allèle/génotype et le lien génotype–phénotype.

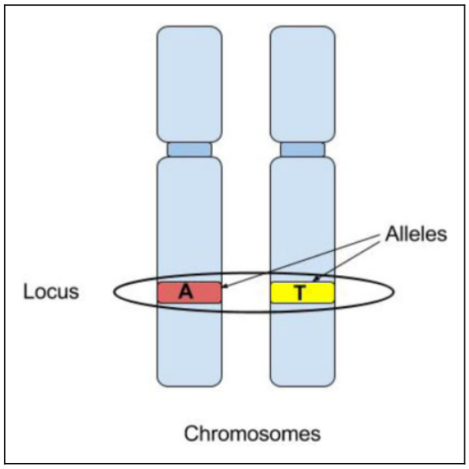

FIGURE 1.2 – Exemple de locus, d'allèle et de génotype. Un locus est une position spécifique sur un chromosome où différentes versions d'un gène, appelées allèles, peuvent exister.

(Wright et Fessele, 2017)

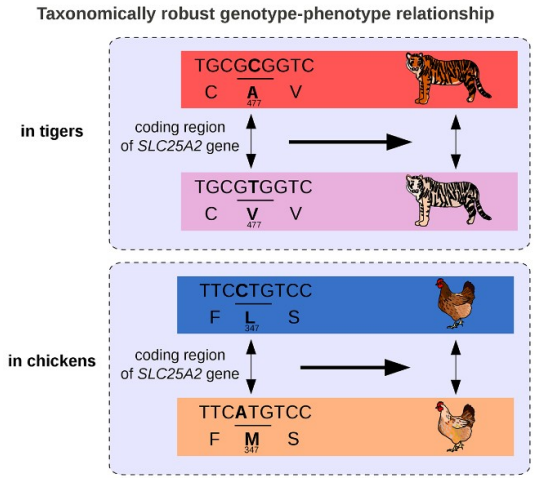

FIGURE 1.3 – Exemple de relation entre génotype et phénotype. Les variations génétiques, telles que les SNPs, peuvent influencer l'expression des gènes et conduire à des différences phénotypiques.

(Orgogozo *et al.*, 2015)

1.2 Données génomiques et vie privée

Le génome humain contient plus de trois milliards de paires de bases réparties sur vingt-trois chromosomes. L'ADN de deux individus diffère en moyenne d'environ 0,5 %, mais cette faible variation peut suffire à révéler des informations sur la santé ou les risques de maladies (Ayday et Humbert, 2017). Les données génomiques peuvent ainsi permettre le diagnostic précoce, les interventions ciblées et révéler des informations sur les membres d'une même famille. Les données génétiques sont à la fois uniques à chaque individu, partagées avec les membres de la famille et inchangées au cours de la vie, ce qui en fait une catégorie particulièrement irrévocable et à forte valeur informative d'un point de vue éthique et en matière de confidentialité. Par exemple, la présence de certaines variantes du gène codant pour l'apolipoprotéine E (ApoE), combinée à des antécédents familiaux, peut augmenter considérablement le risque de développer la maladie d'Alzheimer (Ayday, 2016; Bonomi *et al.*,

2020).

Un autre exemple emblématique est celui du génome d'Henrietta Lacks, une femme décédée en 1951 d'un cancer. Ses cellules, connues sous le nom de cellules HeLa, ont été utilisées à des fins de recherche sans son consentement. Plusieurs années plus tard, les scientifiques ont séquencé l'ADN de ces cellules et ont publié les données sur un site Web public (SNPedia). Cette divulgation a entraîné la fuite d'informations confidentielles sur elle et sa famille, compromettant durablement leur vie privée (Ayday, 2016).

Bien que des efforts d'anonymisation soient généralement appliqués avant le partage des données génomiques, plusieurs études ont démontré qu'ils ne suffisent pas à garantir l'anonymat (Oestreich *et al.*, 2021). En effet, la réidentification d'individus à partir de bases de données ouvertes est rendue possible par le croisement de sources d'informations externes, même sans données personnelles explicites (Gymrek *et al.*, 2013). Plusieurs travaux — notamment Wang *et al.* (2009) (apprentissage d'informations privées à partir de statistiques agrégées) et Wang *et al.* (2017) (exploitation des corrélations entre *SNPs* pour la reconstruction à grande échelle) — montrent que des statistiques agrégées de GWAS peuvent à la fois révéler l'appartenance d'un individu à une cohorte et, en s'appuyant sur la structure de liaison, reconstruire une part substantielle de profils génétiques à partir de jeux statistiques de taille modeste.

C'est pourquoi plusieurs cadres réglementaires, comme le Règlement Général sur la Protection des Données (RGPD) en Europe et la loi *Health Insurance Portability and Accountability Act (HIPAA)* aux États-Unis, s'efforcent de restreindre l'accès à l'utilisation des données génétiques. Cependant, en raison de ces mesures légales, la diffusion de ces informations reste un sujet compliqué et représente un enjeu crucial pour la sécurité (Bonomi *et al.*, 2020; Oestreich *et al.*, 2021). Ces inquiétudes mettent en évidence l'importance d'élaborer des stratégies solides pour assurer la protection des données génomiques. Pour réduire les risques d'exposition, des méthodes comme la protection de la confidentialité différentielle, le chiffrement homomorphe et l'apprentissage fédéré ont été suggérées. Néanmoins, ces solutions présentent encore des limites. En pratique, les mécanismes de défense (régularisation, *Differentially-private stochastic gradient descent (DP-SGD)*, masquage de confiance, distillation, apprentissage fédéré chiffré) visent à réduire le signal d'appartenance au prix d'un com-

promis utilité–confidentialité et, souvent, d'un surcoût computationnel ; une discussion plus étendue est présentée au Chapitre 2 (Oestreich *et al.*, 2021).

1.3 Préliminaires sur l'apprentissage automatique

L'apprentissage automatique a pris une importance croissante en génomique, notamment en raison du volume considérable de données générées par les technologies de séquençage à haut débit. Il est aujourd'hui largement utilisé pour des applications telles que la découverte de médicaments, la prédiction clinique, la médecine personnalisée ou encore l'analyse de l'expression génique, grâce à sa capacité à extraire des modèles à partir de grands ensembles de données complexes et non structurées.

Cependant, le partage de données et de modèles pose des défis en matière de confidentialité. Non seulement la diffusion des données brutes ou statistiques peut porter atteinte à la vie privée, mais le partage des modèles d'apprentissage peut également compromettre la confidentialité des individus inclus dans les ensembles d'entraînement (Shokri *et al.*, 2017; Yeom *et al.*, 2018).

En effet, les modèles d'apprentissage automatique sont susceptibles de mémoriser certaines données spécifiques utilisées lors de l'entraînement, au lieu de se limiter à une généralisation. De plus, ils présentent souvent un comportement différent lorsqu'ils sont exposés à des données vues pendant l'entraînement (membres) par rapport à des données nouvelles (non-membres) (Yeom *et al.*, 2018; Carlini *et al.*, 2021). Ce décalage comportemental fonde les attaques par inférence d'appartenance (MIA), présentées dans le chapitre suivant.

Surapprentissage. On parle de *surapprentissage* lorsque le modèle apprend trop fidèlement les particularités (et le bruit) de D_{train}, au détriment de la *généralisation*. Il se manifeste par un écart marqué entre les performances d'entraînement et de test ainsi que par des sorties très confiantes sur les exemples vus, ce qui accroît la séparabilité membre/non-membre exploitée par les MIA.

Points de données et ensemble de données (contexte génomique).

Dans notre jeu de données de levure, un *point de données* correspond à une souche i identifiée par un

identifiant (par exemple `01_01`, `01_02`, etc.). Cette souche est décrite par un vecteur de génotypes

$$\mathbf{x}_i = (g_{i1}, \ldots, g_{ip}) \in \{-1, 1\}^p,$$

où chaque composante g_{ij} représente le génotype de la souche i au SNP j. Concrètement, les colonnes du fichier de génotypes portent des identifiants tels que `33070 _chrI _33070 _A_T`, etc., et la valeur $g_{ij} \in \{-1, 1\}$ correspond à un codage binaire symétrique de l'allèle observé pour ce SNP chez la souche i (par exemple -1 pour l'allèle de référence et 1 pour l'allèle minoritaire).

Concrètement, les colonnes du fichier de génotypes portent des identifiants tels que `33070_chrI_33070_A_T`, où chaque étiquette encode : (i) la position du SNP sur le chromosome (`33070`), (ii) le chromosome concerné (`chrI`), (iii) la position répétée pour compatibilité avec certains outils génomiques, et (iv) les allèles de référence et alternatif (`A` et `T`). La valeur $g_{ij} \in \{-1, 1\}$ correspond ensuite à un codage binaire symétrique de l'allèle observé chez la souche i pour ce SNP (par exemple -1 pour l'allèle de référence et 1 pour l'allèle alternatif).

La *cible* y_i est un phénotype mesuré pour cette même souche. Dans notre cas, les phénotypes sont des valeurs quantitatives de croissance sous différentes conditions environnementales, organisées en colonnes portant des noms comme `1_CobaltCl`, `1_Xylose_1`, `1_YPD_1`, etc. Pour une tâche de prédiction donnée, on choisit une colonne phénotypique cible (par exemple la croissance sous `1_Xylose_1`) et on note y_i la valeur correspondante pour la souche i.

Un *ensemble de données* supervisé s'écrit alors

$$D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n,$$

où n est le nombre de souches et p le nombre de SNPs (colonnes génotypes). Ce cadre est typiquement en grande dimension ($p \gg n$), avec des corrélations de liaison (LD) entre SNPs.

— **Entraînement** (D_{train}) : on considère $h_\theta(x)$, un algorithme d'apprentissage qui prend x en entrée (vecteur de caractéristiques) et θ comme vecteur de paramètres. L'ensemble des fonctions possibles $\{\forall\theta,\ x \mapsto h_\theta(x)\}$ constitue l'espace des hypothèses. L'objectif est d'ajuster θ en minimisant une fonction de perte empirique $L(\theta; D_{\text{train}})$ afin d'obtenir de bonnes perfor-

mances de *généralisation*. En pratique, le modèle apprend les régularités pertinentes pour la tâche visée (De Cristofaro, 2020).

— **Validation** (D_{val}) : après l'entraînement, on évalue le modèle sur un *ensemble de validation* distinct pour sélectionner les hyperparamètres (p. ex. régularisation, profondeur) et régler des mécanismes comme l'*early stopping*, sans toucher à D_{test} (De Cristofaro, 2020).

— **Évaluation** (D_{test}) : une fois l'architecture et les hyperparamètres figés, on mesure la performance finale sur un *ensemble de test* jamais utilisé aux étapes précédentes, ce qui reflète le comportement attendu en déploiement (prédictions sur des données non vues) (De Cristofaro, 2020).

1.4 Types d'apprentissage automatique

En règle générale, les algorithmes d'apprentissage automatique sont divisés en trois catégories : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement, qui sont déterminés en fonction du type d'information fournie par les données d'entraînement et de diverses tâches d'apprentissage. Au fil des années, de nouvelles catégories, telles que l'apprentissage semi-supervisé, l'apprentissage autosupervisé et l'apprentissage génératif et discriminatif, ont été ajoutées (Rigaki et Garcia, 2023).

1.4.1 L'apprentissage supervisé

Dans l'apprentissage supervisé, les données d'entraînement sont composées d'exemples étiquetés, c'est-à-dire que chaque entrée est associée à une sortie connue. Le modèle apprend à établir une relation entre les entrées et les sorties, ce qui lui permet de prédire correctement l'étiquette d'une nouvelle donnée inconnue. La tâche est appelée classification, si le domaine de sortie est catégoriel. S'il est cardinal, la tâche est régression. Par exemple, le filtrage du pourriel parmi les courriels est une tâche de classification et la prédiction de l'âge est une tâche de régression (Papernot *et al.*, 2018a; De Cristofaro, 2020; Alnuaimi et Albaldawi, 2024; Rigaki et Garcia, 2023). Dans ce mémoire, nous nous intéressons principalement à des tâches de classification supervisée, pour lesquelles plusieurs familles de modèles sont couramment utilisées :

— **Modèles linéaires** (par exemple, régression logistique, *Support Vector Machine (SVM)* linéaire) : ils apprennent une frontière de décision linéaire dans l'espace des caractéristiques et servent souvent de modèles de référence pour évaluer les performances sur des données génomiques (Katsara *et al.*, 2021; Lourenço *et al.*, 2024).
— **Modèles à base d'arbres de décision** (arbres, forêts aléatoires, méthodes d'amplification de gradient comme *eXtreme Gradient Boosting (XGBOOST)*) : ils capturent des relations non linéaires et des interactions entre variables, et sont largement utilisés en pratique pour des tâches de classification tabulaire (Lourenço *et al.*, 2024; Chen et Ishwaran, 2012).
— **Réseaux de neurones profonds** : en particulier les réseaux entièrement connectés et les réseaux convolutifs unidimensionnels, capables de modéliser des relations complexes dans des espaces de grande dimension. Dans ce travail, un réseau convolutionnel 1D est utilisé comme modèle cible pour la prédiction de phénotypes à partir de génotypes (Abdollahi-Arpanahi *et al.*, 2020).

Ces familles de modèles seront réutilisées et discutées dans les chapitres suivants, notamment lors de la présentation des attaques par inférence d'appartenance et de l'état de l'art correspondant.

1.4.2 L'apprentissage non supervisé

Lorsque les entrées ne sont pas étiquetées, on parle d'apprentissage non supervisé. L'objectif est d'identifier des structures sous-jacentes dans les données, en regroupant les observations pour constituer des amas (clusters). Cet apprentissage s'appuie sur des techniques statistiques visant à découvrir des structures latentes ou des régularités cachées au sein de données non étiquetées. On distingue classiquement deux grandes familles : le regroupement (clustering) et l'extraction de règles d'association (Alnuaimi et Albaldawi, 2024; Alzubi *et al.*, 2018).

1.4.3 L'apprentissage semi-supervisé

L'apprentissage semi-supervisé est un mélange de l'apprentissage supervisé et non supervisé. Quand les données étiquetées sont moins nombreuses que celles qui ne le sont pas, cet algorithme est utilisé. D'abord, les données non étiquetées sont utilisées dans l'apprentissage non supervisé afin de les

regrouper. Par la suite, les données étiquetées sont utilisées pour classer les données d'entraînement représentatives de chaque cluster. Cette approche permet d'attribuer automatiquement et à faible coût des étiquettes aux données non étiquetées (Rigaki et Garcia, 2023; Fergus et Chalmers, 2022).

1.4.4 L'apprentissage par renforcement

L'apprentissage par renforcement est une branche particulière de l'apprentissage automatique où un agent apprend à prendre des décisions optimales par essais-erreurs dans un environnement dynamique. Cette méthode ne dépend pas des étiquettes clairement définies par un éducateur, mais d'un mécanisme de récompense. Dans le but de maximiser le cumul des récompenses au fil du temps en développant une stratégie efficace, l'agent reçoit un retour positif lorsqu'il adopte un comportement favorable, sinon un retour négatif (punition). L'apprentissage se fait donc sans connaissance préalable et l'agent commence par des essais aléatoires, puis affine sa stratégie à mesure qu'il accumule de l'expérience. Cette approche est appliquée dans les domaines suivants : la robotique, les jeux vidéo, la conduite autonome (Fergus et Chalmers, 2022; Rigaki et Garcia, 2023; Alnuaimi et Albaldawi, 2024). Ces catégories s'implantent au sein d'architectures de déploiement variées qui conditionnent directement les risques de confidentialité, comme rappelé ci-après.

1.5 Architectures d'apprentissage

1.5.1 Apprentissage centralisé

Les méthodes d'apprentissage centralisé ont tendance à collecter et à stocker les données brutes distribuées générées par divers appareils ou organisations sur un serveur unique ou une grappe de serveurs avec stockage partagé. Dans ce cadre, les données et le modèle sont colocalisés : toutes les données, qu'elles proviennent d'une ou de plusieurs sources, sont regroupées au même endroit pour entraîner un seul modèle. Ce lieu peut être constitué d'une ou de plusieurs machines dans un même centre de données. Cette architecture inclut *MLAAS*, où le propriétaire des données les téléverse sur une plateforme cloud spécialisée (Rigaki et Garcia, 2023). Cette dernière s'occupe ensuite de concevoir et d'optimiser un modèle en fonction des objectifs prédéfinis. Bien que cette solution soit souvent pratique et performante, elle soulève des préoccupations importantes en matière de sécurité et de confidentialité, en particulier dans les contextes sensibles. La transmission sans restriction de données

brutes vers des serveurs tiers et la centralisation dans une région géographique ou une entité unique entraînent une augmentation des risques de violation de la vie privée et de fuites d'informations. De plus, cette approche est confrontée à plusieurs limitations pratiques, telles que la dépendance à la capacité de calcul centralisée, un temps d'apprentissage élevé et l'impossibilité d'accéder à des données distribuées géographiquement sans compromettre leur intégrité ou leur confidentialité (Liu *et al.*, 2019).

1.5.2 Apprentissage fédéré

L'apprentissage fédéré est une approche efficace qui permet d'utiliser des ressources distribuées afin d'entraîner de manière collaborative un modèle d'apprentissage automatique, tout en gardant les données sur chaque appareil ou site. Contrairement aux méthodes centralisées, il ne nécessite pas le transfert des données brutes vers un serveur centralisé. Au lieu de cela, le modèle est entraîné localement sur chaque nœud (appareil ou organisation). Seules les mises à jour du modèle (par exemple, les gradients ou les poids) sont ensuite partagées et agrégées pour former un modèle global. Comme le soulignent McMahan *et al.* (2017) dans leur article fondateur sur l'apprentissage fédéré, cette méthode repose sur le principe fondamental selon lequel il est préférable d'« amener le code vers les données plutôt que d'amener les données vers le code ». Cela répond à des problématiques cruciales concernant la confidentialité, la propriété des données et leur emplacement. L'apprentissage fédéré exploite les ressources de calcul locales réparties dans différentes régions ou institutions. Il s'appuie généralement sur des techniques de protection, telles que le chiffrement ou d'autres mécanismes de défense, pour garantir la sécurité et la confidentialité des données. Cette méthode permet de se conformer aux exigences réglementaires en matière de protection des données tout en exploitant la richesse et la diversité des données distribuées pour construire des modèles plus robustes et généralisables (Liu *et al.*, 2022).

1.6 L'attaque par l'inférence d'appartenance

Puisque les ensembles de données génomiques peuvent contenir des informations sensibles sur les individus, il est essentiel que les modèles d'apprentissage automatique ne révèlent pas, même indirectement, la présence ou l'absence d'un individu dans les données d'entraînement.

L'attaque par inférence d'appartenance est une attaque permettant de prédire si une donnée spécifique est membre ou non d'un ensemble d'entraînement d'un modèle cible (Shokri *et al.*, 2017; Hu *et al.*, 2022). Cette attaque repose sur le fait que les modèles se comportent souvent différemment lorsqu'ils traitent des données vues pendant l'entraînement (membres) comparées à des données inconnues (non-membres).

L'attaquant peut avoir deux niveaux de connaissance : si l'attaquant possède toutes les informations sur le modèle cible, y compris sa distribution de données d'entraînement, son architecture et ses paramètres, l'attaque est qualifiée de « boîte blanche ». Dans le cas d'une attaque en boîte blanche, l'adversaire a accès aux gradients et aux poids internes du modèle, ce qui lui permet de reconstruire des informations précises sur les échantillons d'entraînement. Ces attaques sont donc plus efficaces et exigent des mesures de défense plus solides.

En revanche, si l'attaquant ne dispose que d'informations limitées sur la distribution des données d'entraînement et effectue des requêtes sur le modèle cible sans avoir accès à ses paramètres internes, l'attaque est qualifiée de « boîte noire ». Dans les MIAs basées sur un classificateur binaire, le modèle d'attaque est un classificateur binaire qui déduit les membres et les non-membres de l'ensemble des données d'entraînement du modèle cible. Pour ce faire, l'approche du modèle d'ombre, présentée par Shokri *et al.* (2017), est largement utilisée. Dans cette technique, l'attaquant crée un ou plusieurs modèles semblables au modèle cible, entraînés sur des jeux de données artificiels reproduisant sa distribution, afin de simuler son comportement. En comparant les réponses obtenues, il est alors possible de distinguer les membres des non-membres de l'ensemble d'entraînement. En boîte blanche, le modèle d'ombre est construit avec la même structure et le même algorithme d'apprentissage que ceux du modèle cible. En boîte noire, l'attaquant obtient le vecteur de prédiction d'un enregistrement d'entrée uniquement lorsqu'il interroge le modèle cible.

En ce qui concerne les modèles d'ombre, l'attaquant a accès à la fois aux données d'entraînement et aux données de test. Cela lui permet de créer un ensemble de données qui contient les caractéristiques et la vérité de terrain de l'appartenance des enregistrements de données d'entraînement et de test. En utilisant cette base de données, l'attaquant peut entraîner un modèle d'attaque fondé sur un algorithme de classification binaire.

Variantes d'attaque. Au-delà des modèles d'ombre, les MIA se déclinent en approches (i) *fondées sur le score de confiance* (vecteur de probabilités, top-k, entropie), (ii) *fondées sur la perte* (seuil sur $\ell(x, y)$ ou rapport de vraisemblance calibré, p. ex. LiRA), (iii) *label-only* (sans accès aux probabilités, via agrégation d'augmentations et marge de décision), et (iv) *white-box* (gradients/poids). Ces familles diffèrent par les hypothèses d'accès et le signal exploité, mais partagent le même objectif : discriminer membres et non-membres.

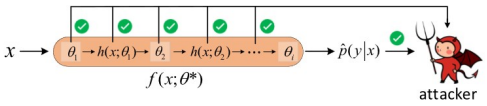

FIGURE 1.4 – Dans ce scénario, l'adversaire bénéficie d'un accès complet à l'architecture, aux poids et aux gradients du modèle cible, ce qui facilite la mise en œuvre d'attaques très précises. Cependant, ce type d'attaque repose sur une hypothèse souvent irréaliste dans les applications réelles (Hu *et al.*, 2022).

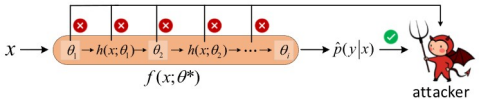

FIGURE 1.5 – Dans ce cas, l'adversaire n'a accès qu'aux sorties du modèle cible (par exemple, les scores de prédiction), sans aucune information sur sa structure interne. Cela rend l'attaque plus difficile à concevoir, mais aussi plus générique et réaliste, notamment dans les contextes de *MLAAS*. C'est ce défi que ce mémoire cherche à relever (Hu *et al.*, 2022).

Note (notation commune aux Figures 1.4–1.5). x : entrée ; $h^{(\ell)}$: activation de la couche ℓ ; $f(x;\theta)$: modèle paramétré par θ ; θ^* : paramètres appris ; $p(y \mid x)$: distribution des probabilités ; $\hat{y}$: prédiction ; *attacker* : adversaire.

Protocole d'évaluation. Pour évaluer l'efficacité d'une attaque par inférence d'appartenance, il est nécessaire de constituer un ensemble de données d'évaluation dont le statut d'appartenance est connu de manière contrôlée. Dans notre protocole, un échantillon est considéré comme *membre* s'il provient du jeu d'entraînement du modèle cible, et comme *non-membre* s'il appartient à un sous-ensemble dédié de données jamais utilisées pendant l'entraînement (jeu *unseen*). Ces deux groupes sont construits de manière équilibrée afin d'éviter un biais lié aux proportions de classes. Pour chaque échantillon, nous collectons uniquement la sortie du modèle cible (probabilité prédite), sans accès à sa structure interne, conformément au cadre en boîte noire. Les paires $(\text{score}, \text{étiquette})$ ainsi obtenues constituent le jeu d'évaluation de l'attaque, sur lequel nous mesurons les métriques classiques des MIA : exactitude, précision, rappel, F1-score, ainsi que le couple TPR/FPR et la courbe ROC. Ce protocole, appliqué

de manière identique aux deux méthodologies proposées, garantit une comparaison cohérente et une évaluation contrôlée de la capacité du modèle d'attaque à distinguer membres et non-membres. Les détails complets sont fournis au Chapitre 2.

1.7 Objectif et contributions du projet

L'objectif de cette étude est d'analyser et d'évaluer la vulnérabilité des modèles d'apprentissage automatique face aux attaques qui révèlent la confidentialité des données, comme l'attaque par inférence d'appartenance. Une étude antérieure menée par Chen *et al.* (2020) a exploré la MIA en boîte blanche sur des données génomiques. Les auteurs ont démontré que, même dans ce contexte, des techniques de protection comme la confidentialité différentielle peuvent atténuer le risque de réidentification. Cependant, le scénario boîte blanche suppose un accès total au modèle, ce qui est rarement le cas dans les environnements réels.

Contrairement à cette approche, notre travail se concentre sur un scénario en boîte noire, plus représentatif des usages réels (par exemple dans les services *MLAAS*), où l'adversaire ne dispose que des sorties du modèle cible. Plus précisément, nous cherchons à implémenter une attaque par inférence d'appartenance contre un modèle prédictif inférant un phénotype à partir de données génomiques. Notre objectif est de développer un modèle d'attaque généralisable, capable de s'adapter à différentes configurations sans nécessiter une connaissance fine du modèle attaqué.

Enfin, nous visons à évaluer, dans un cadre réaliste de boîte noire, la capacité d'un adversaire à inférer l'appartenance d'échantillons à un modèle génomique prédictif et à concevoir un modèle d'attaque généralisable limitant les faux positifs. Pour ce faire, nous passons en revue les attaques contre les modèles d'apprentissage automatique avec un focus sur les MIA, synthétisons les approches récentes et leurs métriques d'évaluation, concevons et mettons en œuvre une MIA en boîte noire sur données génomiques, puis menons une analyse expérimentale démontrant la capacité du modèle à distinguer de façon fiable membres et non-membres tout en maîtrisant les biais.

1.8 Conclusion

L'exploitation des données génomiques par l'apprentissage automatique ouvre des perspectives prometteuses en biologie et en médecine. Cependant, elle s'accompagne de risques importants en matière de confidentialité, notamment liés à la possibilité d'identifier des individus ou d'inférer des informations sensibles à partir de leurs données.

En effet, toute fuite d'information génomique peut avoir des conséquences durables non seulement pour l'individu concerné, mais aussi pour sa famille. Ce chapitre a mis en lumière l'importance des données génomiques, les risques liés à leur exposition, ainsi que les vulnérabilités spécifiques des modèles d'apprentissage automatique face aux attaques visant à révéler des informations confidentielles.

Parmi ces menaces, l'attaque par inférence d'appartenance (MIA) constitue un risque particulièrement préoccupant, car elle permet à un adversaire de déterminer si un échantillon a été utilisé pour entraîner un modèle donné. Ce type d'attaque est d'autant plus redoutable qu'il peut s'appliquer dans des scénarios réalistes de boîte noire, où l'adversaire ne connaît ni les données ni la structure du modèle.

Dans ce mémoire, nous proposons une attaque MIA en boîte noire appliquée à un modèle prédictif entraîné sur des données génomiques de levure. Notre objectif est d'évaluer la faisabilité et la généralisation de ce type d'attaque, en mettant l'accent sur la robustesse et la précision du modèle d'attaque. Le chapitre suivant présente l'état de l'art en matière de protection de la vie privée en apprentissage automatique, en détaillant les différentes formes d'attaques existantes ainsi que les stratégies de défense actuellement proposées.

CHAPITRE 2
PRÉSENTATION DE L'ÉTAT DE L'ART DE L'ATTAQUE D'INFÉRENCE D'APPARTENANCE

Après avoir terminé le projet du génome humain (*Human genome project (HGP)*), le perfectionnement des techniques de séquençage ainsi que l'essor des domaines de l'informatique et des télécommunications ont permis d'accumuler, de classer, d'analyser et de diffuser une immense quantité de données génétiques. L'accessibilité croissante des données génomiques et leur nature sensible ont suscité d'importantes préoccupations en matière de confidentialité. Étant donné que les modèles d'apprentissage automatique (ML) peuvent fonctionner efficacement avec de vastes ensembles de données et fournir des prédictions précises, leur utilisation en biologie, et plus particulièrement dans l'analyse des données génomiques, est devenue de plus en plus populaire. L'intégration des modèles d'apprentissage automatique dans l'analyse des données génomiques a apporté des avancées considérables. Elle a notamment permis des progrès en médecine personnalisée, en détection précoce des maladies et en recherche biologique. Cependant, entraîner les modèles d'apprentissage automatique sur des ensembles de données sensibles pose des risques significatifs de fuite d'informations, car ces modèles peuvent mémoriser et exposer certaines caractéristiques des données d'entraînement. Ces vulnérabilités permettent d'attaquer les modèles d'apprentissage automatique afin de divulguer des informations sensibles sur la confidentialité des données d'entraînement de l'apprentissage automatique (Hu *et al.*, 2022).

Dans la suite de ce chapitre, nous (i) organisons un panorama des principales attaques contre les modèles d'apprentissage automatique, en distinguant celles qui visent la sécurité du modèle de celles qui ciblent la vie privée des données, (ii) formalisons les attaques par inférence d'appartenance ainsi que les modèles d'adversaire et les niveaux de sortie considérés, (iii) présentons et comparons les principales approches d'attaque (modèles d'ombre, heuristiques sur les scores, comparaison différentielle), (iv) passons en revue l'état de l'art des MIA sur données génomiques et biomédicales, et (v) synthétisons les stratégies de défense existantes (masquage de la confiance, régularisation, confidentialité différentielle, distillation des connaissances) en les replaçant dans le contexte de ce mémoire.

2.1 Rappels sur l'apprentissage automatique

L'apprentissage automatique est une branche de l'intelligence artificielle qui permet à un système informatique d'apprendre à partir de données et d'améliorer ses performances sans être explicitement programmé pour chaque tâche. Il repose sur des algorithmes capables de détecter des motifs, de faire des prédictions et de prendre des décisions dans des domaines variés tels que la santé, la finance, la sécurité, la reconnaissance d'image ou encore la biologie computationnelle (Alnuaimi et Albaldawi, 2024; Muhamedyev, 2015). Nous renvoyons le lecteur au chapitre 1 pour une introduction générale plus détaillée à l'apprentissage automatique.

2.2 Attaques sur les modèles d'apprentissage automatique

En raison du développement de l'intelligence artificielle et de l'intégration croissante des modèles d'apprentissage automatique dans divers domaines, comme la santé, la finance ou la sécurité et la justice, la sécurité des modèles d'apprentissage automatique et la protection des données sensibles sont devenues un sujet crucial. Des recherches récentes ont révélé que les modèles d'apprentissage automatique sont exposés à diverses attaques. Ces attaques peuvent avoir comme objectif de manipuler le comportement du modèle, d'en extraire des informations sensibles, ou encore de compromettre la confidentialité des données d'entraînement. Les systèmes d'apprentissage automatique sont confrontés à une variété de menaces, qui peuvent survenir à différentes étapes de leur cycle de vie, soit au moment de l'entraînement, soit durant la phase d'inférence. Ces menaces exploitent les vulnérabilités des modèles, des données ou des interfaces d'accès (Xue *et al.*, 2020).

Comme illustré à la figure 2.1, un attaquant peut intervenir :

- **Phase d'entraînement :**
  - Empoisonnement des données : manipulation des exemples/étiquettes pour biaiser l'apprentissage.
  - Insertion d'échantillons malveillants : ajout de points conçus pour dégrader le modèle.
- **Phase d'inférence :**
  - Attaques adversariales : envoi d'entrées soigneusement construites pour provoquer des erreurs.

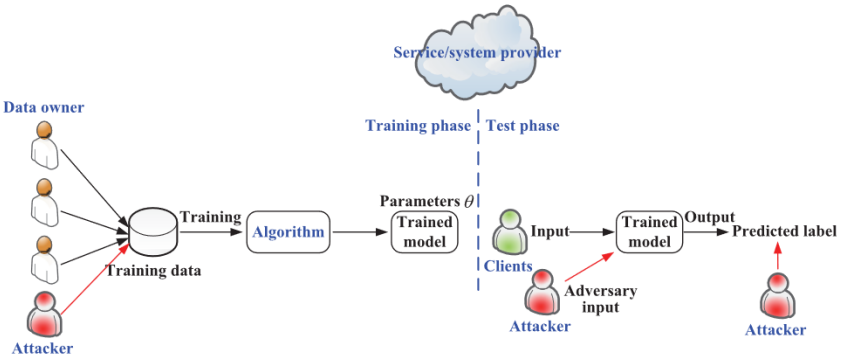

FIGURE 2.1 – Les surfaces d'attaque dans un pipeline d'apprentissage automatique. Les attaquants peuvent agir pendant l'entraînement (empoisonnement de données) ou pendant l'inférence (attaques adversariales, inférence d'appartenance, etc.).

— Inférence d'appartenance (MIA) : déduire si un exemple a servi à l'entraînement.
— Extraction de modèle : répliquer le comportement (ou l'interface de programmation applicative, (*Application programming interface (API)*)) du modèle cible.

2.2.1 Attaques visant la sécurité du modèle

2.2.1.1 L'attaque adversariale

Les attaques adversariales sont les attaques les plus courantes dans le domaine de l'apprentissage automatique. Dans ce genre d'attaque, en ajoutant une petite perturbation aux données, le modèle de classification se trompe. L'attaque adversariale peut être soit ciblée, soit non ciblée. Dans l'attaque adversariale ciblée, les données sont changées pour obliger le modèle à prédire un résultat particulier. En revanche, l'attaque adversariale non ciblée ne cherche pas à obliger le modèle à une sortie particulière, mais cherche simplement à entraîner une quelconque mauvaise prédiction du modèle (Rahman *et al.*, 2023). Ce type d'attaque est réalisé pendant la phase d'inférence, une fois le modèle entraîné, afin de manipuler ses prédictions sans en modifier les paramètres internes.

2.2.1.2 L'attaque par empoisonnement

L'empoisonnement de données est une méthode d'attaque visant la phase d'entraînement d'un modèle d'apprentissage automatique. L'adversaire introduit dans l'ensemble d'entraînement des données malicieusement conçues, qu'on appelle des échantillons empoisonnés. Ces données semblent normales à première vue, mais elles ont été soigneusement manipulées pour influencer négativement l'apprentissage du modèle. Le but peut être de faire échouer totalement l'entraînement, de réduire les performances globales du modèle, ou encore de créer des comportements erronés ciblés sur certains types d'entrées. Ce type d'attaque est particulièrement insidieux, car l'attaquant ne modifie pas directement le fonctionnement du modèle ; il respecte les étapes classiques du processus d'apprentissage, mais agit uniquement sur les données fournies. En s'appuyant sur la confiance accordée aux données d'entrée, l'adversaire peut corrompre subtilement le modèle, parfois sans laisser de trace visible (Tian *et al.*, 2022).

2.2.2 Attaques visant la vie privée des données

2.2.2.1 L'attaque par inversion de modèle

Dans cette attaque, l'adversaire essaie d'extraire des informations des données d'entraînement du modèle. L'adversaire peut utiliser la sortie du modèle afin de reconstruire les données d'entrée pour induire en erreur le modèle cible. Fredrikson *et al.* (2015) ont été les premiers à proposer une méthode pour reconstituer les caractéristiques personnelles d'un individu, telles que son apparence faciale ou son profil génétique, à partir des scores de confiance fournis par un algorithme d'apprentissage automatique. Leur étude a montré qu'un attaquant peut exploiter les prédictions du modèle pour inverser son comportement et générer une estimation plausible de l'entrée d'origine, même dans une situation de boîte noire. Par exemple, dans le cas d'un système de reconnaissance faciale, l'attaque permettrait de reconstruire l'image d'un visage à partir de simples sorties de probabilité du modèle. Cette attaque met en évidence une fuite potentielle d'informations, même lorsque les données brutes ne sont pas directement divulguées. Elle s'effectue généralement pendant la phase d'inférence, lorsque le modèle est déjà entraîné et accessible (Fredrikson *et al.*, 2015; Rahman *et al.*, 2023).

2.2.2.2 L'attaque par extraction du modèle

Dans l'attaque par extraction de modèles, l'adversaire interroge le modèle cible en passant par une interface de prédiction (comme une API) afin de reconstruire un modèle équivalent. Bien qu'il ne possède aucune connaissance préalable sur l'architecture ou les paramètres internes du modèle cible, il peut choisir des entrées et observer les sorties correspondantes. À partir de ces paires entrée-sortie, l'attaquant entraîne un modèle substitut qui imite le comportement du modèle original (Zhang *et al.*, 2021). Cette attaque se déroule typiquement pendant la phase d'inférence, une fois le modèle entraîné et exposé au travers d'une interface de requête, comme c'est souvent le cas dans les services *MLAAS*. Ce procédé ne sert pas seulement à reproduire les performances du modèle cible ; il permet aussi d'effectuer d'autres attaques, telles que des attaques adversariales ou des attaques d'inférence, tout en présentant un risque élevé de vol de propriété intellectuelle. Ce type d'attaque a été démontré de manière concrète par Tramèr *et al.* (2016) dans leur étude sur les modèles accessibles à travers des API dans les services *MLAAS*, illustrant ainsi la facilité avec laquelle un adversaire peut extraire un modèle complexe en boîte noire (Tramèr *et al.*, 2016).

2.2.2.3 L'attaque par l'inférence d'appartenance

L'attaque par inférence d'appartenance est une menace sérieuse pour la confidentialité des données, en particulier dans les domaines sensibles, comme la santé ou la génomique. Un adversaire tente de déterminer si un échantillon de données a été utilisé lors de l'entraînement d'un modèle d'apprentissage automatique. Cette capacité à inférer l'appartenance peut entraîner la divulgation de données personnelles sensibles. Cette attaque se déroule typiquement pendant la phase d'inférence, une fois le modèle entraîné et accessible, et elle peut être mise en œuvre même avec un accès limité à l'information sur le modèle cible (Shokri *et al.*, 2017).

2.2.3 Détails sur les attaques par inférence d'appartenance

L'attaque par inférence d'appartenance détermine si un individu est dans l'ensemble d'entraînement d'un modèle ciblé ou non. Cette attaque montre une menace majeure pour la confidentialité des données, surtout dans les domaines sensibles comme la santé et la génomique.

Imaginez un hôpital qui utilise un modèle d'apprentissage automatique sur le service infonuagique afin de diagnostiquer une maladie à partir des données génomiques. Ce modèle a déjà été entraîné avec les données des patients ayant fourni leurs données génomiques. Un adversaire peut interroger ce modèle pour identifier si le génome d'une personne a servi à entraîner le modèle ou non, ce qui pourrait révéler des informations sensibles sur son état de santé.

2.2.4 Pourquoi l'attaque par l'inférence d'appartenance fonctionne-t-elle ?

Quand les modèles d'apprentissage automatique sont parfaitement ajustés sur les données d'entraînement, mais qu'ils généralisent mal aux données de test, on parle de surapprentissage. La relation entre le surapprentissage de modèle et la force de l'attaque par inférence d'appartenance en boîte noire est montrée par Shokri *et al.* (2017). Ensuite, Yeom *et al.* (2018) a confirmé cet effet de surapprentissage. L'adversaire utilise ce comportement du modèle afin de déterminer si un enregistrement a servi à entraîner le modèle ou non. Le surapprentissage est causé par la complexité du modèle et la taille limitée de l'ensemble de données d'entraînement (Hu *et al.*, 2022). Les modèles complexes, tels que les réseaux de neurones profonds (*Deep neural network (DNN)*), qui comportent un grand nombre d'hyperparamètres, peuvent mémoriser en détail les données d'entraînement, surtout lorsqu'ils sont entraînés pendant plusieurs époques. Par ailleurs, lorsqu'on dispose d'un ensemble de données d'entraînement de taille limitée, le modèle a plus de difficulté à refléter correctement la diversité réelle des données. Cela réduit sa capacité à bien s'adapter à de nouvelles situations. De plus, le type de modèle cible joue un rôle important dans la réussite d'une attaque par inférence d'appartenance. Lorsque la limite de décision du modèle n'est pas facilement influencée par les données spécifiques, le modèle est généralement plus résistant à cette forme d'attaque. Hu *et al.* (2022) ont montré que certains modèles, tels que le classifieur bayésien naïf, sont moins vulnérables aux attaques par inférence d'appartenance à cause de leur fonctionnement probabiliste et de leur tendance limitée à mémoriser les données d'entraînement. Par contre, les arbres de décision ou les réseaux de neurones, qui sont des modèles plus complexes, ont tendance à apprendre des schémas spécifiques à partir des exemples vus pendant l'entraînement. Cela peut entraîner des fuites d'informations. Les auteurs soulignent également que le nombre de classes dans un jeu de données a un impact non négligeable sur l'efficacité de l'attaque : plus ce nombre est élevé, plus le modèle risque d'avoir un comportement différencié selon les entrées, ce qui facilite la détection des données ayant servi à l'apprentissage (Truex *et al.*, 2019).

2.2.5 Niveau de connaissance de l'adversaire

L'efficacité d'une attaque par inférence d'appartenance dépend des informations dont dispose l'adversaire sur (i) le modèle cible et (ii) les données d'entraînement (Hu *et al.*, 2022).

La connaissance du modèle cible correspond au niveau d'accès de l'adversaire à l'architecture, à l'algorithme d'apprentissage et, dans certains cas, aux paramètres internes du modèle. Cette connaissance donne lieu à deux cadres, soit l'attaque en boîte blanche et l'attaque en boîte noire. Dans un cadre de boîte blanche, l'adversaire a un accès complet à ces informations, ce qui permet de concevoir des attaques très précises. En revanche, dans un cadre de boîte noire, l'adversaire n'a accès qu'aux prédictions du modèle (scores de confiance, classes ou logits) sans connaissance de son fonctionnement interne.

En parallèle, l'adversaire peut également avoir une connaissance partielle ou complète de la distribution des données d'entraînement. Il est souvent supposé que l'adversaire peut obtenir un jeu de données d'ombre provenant de la même distribution que les données d'entraînement.

Il est crucial de noter que, dans les attaques en boîte noire, le degré de connaissance peut varier selon les informations fournies par le vecteur de prédiction. En effet, les attaques d'inférence d'appartenance en boîte noire dépendent du niveau de sortie du modèle cible. Il existe trois catégories selon le niveau d'information renvoyé par le modèle :

- Vecteur complet de probabilités : l'adversaire infère l'étiquette et peut calculer des mesures comme la perte (p. ex., entropie croisée). *C'est le cas le plus riche en information et, en général, le plus favorable aux attaques.*
- Top-k probabilités : l'information est réduite, mais l'adversaire peut encore construire un modèle d'attaque en exploitant des motifs partiels (rang, écarts entre scores, etc.), avec des performances typiquement inférieures au cas précédent.
- L'étiquette seule (Label-only) : même lorsque seule l'étiquette prédite est fournie, des attaques restent possibles (p. ex., via des signaux de décision). *Ce n'est pas le scénario le plus performant, mais il demeure préoccupant car il montre que masquer les scores ne suffit pas.*

2.2.6 Niveau d'approche de l'attaque d'inférence d'appartenance

Selon la stratégie de l'attaquant et des ressources disponibles, différentes approches ont été développées pour mener des attaques par inférence d'appartenance. Ces approches se distinguent principalement par la manière dont elles exploitent le modèle cible, la quantité de connaissances nécessaires ainsi que les outils employés pour estimer la probabilité d'appartenance d'une donnée à l'ensemble d'entraînement. Voici un aperçu des trois principales familles d'approches :

- Approche par classificateur (modèle d'attaque supervisé) : cette technique s'appuie sur la formation d'un modèle d'attaque, généralement un classificateur binaire, qui prédit si une donnée a été vue ou non par le modèle cible. Pour ce faire, l'attaquant crée un ou plusieurs modèles d'ombre, chacun entraîné sur un ensemble de données simulant la distribution du modèle cible. Le modèle d'attaque est ensuite entraîné à partir des réponses du modèle d'ombre sur des exemples connus comme étant membres (ensemble d'entraînement) ou non (ensemble de tests). Il apprend ainsi à détecter des différences de comportement du modèle sur ces deux types d'exemples. Cette méthode demande un minimum de connaissance du domaine ou de la distribution des données d'entraînement, mais elle fonctionne bien, même dans un cadre en boîte noire. Cette approche est illustrée notamment par l'attaque pionnière de Shokri *et al.* (2017).
- Approche basée sur des métriques heuristiques : dans cette famille d'attaques, l'adversaire ne construit pas explicitement de modèle d'attaque. Il exploite directement des mesures simples dérivées des sorties du modèle cible pour estimer la probabilité d'appartenance d'un échantillon. Les métriques couramment utilisées sont les suivantes :
  - **La perte (loss)** : on suppose que les membres de l'ensemble d'entraînement ont tendance à générer une perte plus faible.
  - **La confiance maximale** : la plus haute probabilité attribuée à une classe.
  - **L'entropie de la prédiction** : mesure l'incertitude du modèle.
- Approche par comparaison différentielle (analyse statistique) : ici, l'attaquant utilise des méthodes statistiques pour comparer le comportement du modèle sur une donnée cible avec celui observé sur un ensemble de données considérées comme non-membres. L'objectif est de formuler l'appartenance comme une hypothèse statistique à tester. Cette méthode est particulièrement utile pour classer les enregistrements les plus vulnérables à l'inférence. L'attaque

pragmatique de (Long *et al.*, 2020) illustre cette méthode avec l'utilisation de tests d'hypothèses et de valeurs-p (p-values) pour détecter les enregistrements les plus vulnérables.

Plus concrètement, supposons que l'attaquant dispose d'un ensemble de référence de $m = 1000$ individus dont il sait qu'ils ne font pas partie de l'ensemble d'entraînement (non-membres). Il interroge le modèle sur ces 1000 non-membres et obtient une distribution de pertes $\{\ell(x_j, y_j)\}_{j=1}^{m}$, typiquement comprises entre, par exemple, $0{,}5$ et $0{,}8$. Pour une donnée cible $(x^\star, y^\star)$, il calcule la perte $\ell(x^\star, y^\star)$ et obtient une valeur très faible, par exemple $0{,}05$. En comptant combien de non-membres ont une perte inférieure ou égale à cette valeur, il obtient, disons, 5 individus, soit une proportion

$$p = \frac{5}{1000} = 0{,}005.$$

Cette valeur-p empirique est très faible sous l'hypothèse « non-membre » (par exemple $p < 0{,}01$) ; l'attaquant en déduit alors que le comportement du modèle sur $(x^\star, y^\star)$ ressemble beaucoup plus à celui observé sur des exemples d'entraînement, et conclut que cet enregistrement est probablement un membre de l'ensemble d'entraînement.

TABLE 2.1 – Comparaison des approches d'attaque par inférence d'appartenance

Approche	**Principe**	**Exemple**	**Avantages / Limites**
Par classificateur	Entraîner un modèle d'attaque (ex. binaire) basé sur les sorties de modèles d'ombre simulant le modèle cible	Shokri *et al.* (2017)	Haute précision, nécessite beaucoup de données similaires au modèle cible.
Basée sur des métriques	Utilise des scores comme la perte, la confiance, l'entropie pour détecter l'appartenance sans apprentissage explicite	Salem *et al.* (2019); Yeom *et al.* (2018)	Facile à implémenter, mais souvent moins performant.
Comparaison différentielle	Applique des tests statistiques pour détecter des écarts de comportement entre membres et non-membres	Long *et al.* (2020)	Ne nécessite pas de données d'entraînement, mais est sensible aux variations naturelles.

2.3 État de l'art sur les attaques par inférence d'appartenance

Tout d'abord, Homer et collaborateurs ont prouvé qu'un attaquant peut exploiter les statistiques publiées sur un jeu de données génomiques pour inférer la présence d'un individu donné. Leur étude met en évidence qu'il est possible de déduire la participation d'une personne (ou d'un proche) à une étude, même lorsque les données génétiques individuelles ne sont pas directement divulguées, mais uniquement des résumés statistiques, tels que les fréquences alléliques ou les distributions de génotypes. En s'appuyant sur une approche basée sur des microréseaux d'ADN à haute densité pour le génotypage

des SNP, ils ont proposé un cadre théorique pour comparer les fréquences alléliques d'un mélange d'ADN avec celles d'une population de référence et d'un individu spécifique. Grâce à une mesure de distance spécifique et à un test statistique, leur méthode permet de détecter la présence d'un individu dans un mélange complexe d'ADN, même lorsqu'il ne contribue qu'à une infime proportion (moins de 0,1%) (Homer *et al.*, 2008; Hu *et al.*, 2022).

Les attaques par inférence d'appartenance ont été introduites de manière marquante par Shokri *et al.* (2017). Ils ont démontré qu'il est possible pour un attaquant de divulguer les données d'entraînement du modèle d'apprentissage automatique, même dans le cadre d'une boîte noire. Pour atteindre cet objectif, ils mettent en œuvre une méthode du modèle d'ombre. Le but du modèle d'ombre est de trouver les liens entre les données et les étiquettes lorsque l'attaquant obtient les sorties du modèle cible en lui fournissant les entrées. Pour mettre en œuvre cette méthode, l'attaquant crée n modèles d'ombre. Chaque modèle d'ombre est entraîné sur un ensemble de données distinct de celui utilisé pour le modèle cible, mais issu de la même distribution que l'ensemble de données d'entraînement.

Le modèle d'ombre doit se former de la même façon que le modèle cible, mais dans le cadre de boîte noire, l'attaquant n'a aucune connaissance à propos de la structure du modèle et des paramètres. Une fois les modèles d'ombre entraînés, l'attaquant les interroge à l'aide d'exemples connus (membres) et inconnus (non-membres) pour obtenir les sorties correspondantes. Ces données servent alors à entraîner un modèle d'attaque capable de prédire, pour une nouvelle entrée, si celle-ci a été utilisée ou non dans l'entraînement du modèle cible. Plus le nombre de modèles d'ombre est élevé, plus le modèle d'attaque sera précis, car il aura été exposé à une plus grande diversité de comportements issus de modèles similaires au modèle cible.

Enfin, toutes les sorties sont utilisées afin d'entraîner un modèle classificateur, appelé modèle d'attaque, qui apprend à distinguer si une donnée a été vue (membre) ou non vue (non-membre) par le modèle cible, en se basant uniquement sur les réponses fournies par celui-ci. Ce modèle peut ensuite servir à faire des prédictions d'appartenance sur de nouvelles données. Il révèle ainsi des informations sensibles sur l'ensemble d'entraînement du modèle cible (Shokri *et al.*, 2017).

Backes et collaborateurs ont exploré une nouvelle dimension de la confidentialité des données biomé-

dicales en analysant les attaques par inférence d'appartenance dans le contexte des études basées sur les expressions de microARN (miARN), de petites molécules d'ARN non codantes qui régulent l'expression des gènes en modulant la traduction ou la dégradation des ARN messagers. Contrairement aux données génétiques statiques, comme le génome, les expressions de miARN sont influencées de manière dynamique par l'état de santé d'un individu, ce qui en fait des biomarqueurs puissants, mais sensibles.

Selon les chercheurs, il est possible pour un adversaire d'estimer avec une grande précision la participation d'une personne spécifique à une étude, même si seules des données statistiques globales, telles que la moyenne d'expression, sont divulguées. En utilisant des données publiques sur les miRNA, ils montrent que, dans les jeux de données associés à des maladies, les attaques peuvent atteindre un taux de vrais positifs de 77%, avec moins de 1% de faux négatifs.

Pour cela, deux approches d'attaque ont été proposées. La première consiste à calculer la distance L1, une mesure mathématique qui calcule la somme des différences absolues entre les niveaux d'expression du miARN d'un individu cible et les moyennes issues des données de l'étude. Plus cette distance est faible, plus cela suggère que le profil de l'individu est compatible avec celui des participants. La seconde méthode repose sur un test du rapport de vraisemblance (likelihood ratio test), une technique statistique qui compare la probabilité qu'un individu appartienne au groupe étudié et celle qu'il en soit exclu. Ce test évalue la compatibilité d'un profil à deux suppositions : la première est que la personne étudiée est présente (hypothèse alternative) ; la seconde, qu'elle n'est pas là (hypothèse nulle). Parmi les deux méthodes, la seconde s'est révélée la plus efficace.

Devant ces inquiétantes observations, les auteurs ont proposé des mesures pratiques pour renforcer la sécurité des participants. Ils ont conseillé de ne pas publier de statistiques agrégées si le jeu de données contient moins de quelques centaines d'individus, ce qui complique l'identification d'une personne en particulier. Pour les ensembles de données de plus petite taille, ils ont suggéré d'ajouter une part d'aléa aux résultats publiés grâce à un procédé statistique qui introduit une incertitude maîtrisée. Cette méthode permet de brouiller suffisamment les informations, tout en conservant une utilité minimale pour l'analyse scientifique. Une autre méthode recommandée consiste à réduire significativement le nombre de statistiques divulguées, ce qui permet de limiter les risques de réidentification (Backes

et al., 2016).

Ensuite, Liu et collaborateurs ont proposé une nouvelle approche, "SocInf", une attaque d'inférence d'appartenance en boîte noire, sans connaître l'architecture du modèle cible ni les données d'entraînement. Cette méthode a été évaluée à partir de données de santé provenant de réseaux sociaux. Son principe de base consiste à développer un modèle de mimétisme qui imite le comportement du modèle cible. Pour ce faire, l'attaquant crée des données synthétiques similaires en format aux données d'origine. Il les regroupe ensuite en fonction des prédictions obtenues. Un processus d'apprentissage est ensuite mis en œuvre pour entraîner le modèle de mimétisme jusqu'à ce que ses sorties deviennent difficiles à distinguer de celles du modèle cible. Sur cette base, un modèle d'attaque est ensuite appris pour estimer si une donnée particulière a été utilisée lors de l'entraînement initial.

Cette approche est particulièrement utile dans ce domaine, car elle démontre qu'un attaquant peut atteindre une haute précision d'inférence sans avoir besoin de détails sur le modèle cible, contrairement à des méthodes plus traditionnelles, comme l'entraînement de modèles d'ombre. Selon SocInf, il suffit de pouvoir consulter les prédictions du modèle pour mettre en évidence des faiblesses importantes, surtout si le modèle montre un surapprentissage par rapport à ses données d'entraînement. Même dans des situations réelles et limitées, comme les services d'intelligence artificielle disponibles en ligne, un adversaire peut détecter si une personne fait partie de l'ensemble de données d'entraînement, ce qui représente une grave menace pour la confidentialité, en particulier dans les domaines sensibles tels que les données médicales ou génétiques (Liu *et al.*, 2019). En outre, Salem et collaborateurs ont démontré qu'un attaquant peut identifier si un point de données particulier a été utilisé pour entraîner le modèle cible, même s'il utilise un jeu de données distinct. Ils ont proposé une technique appelée « attaque par transfert de données », qui assouplit les hypothèses traditionnelles des attaques par inférence d'appartenance. Contrairement à l'attaque de Shokri *et al.* (2017), qui supposait la disponibilité de plusieurs modèles d'ombre et d'un jeu de données issu de la même distribution que celui du modèle cible, Salem et collaborateurs ont montré qu'il est possible de mener une attaque réussie avec un seul modèle d'ombre, voire avec une architecture différente de celle du modèle cible.

En effet, le modèle d'attaque parvient à repérer des différences générales dans la manière dont le modèle cible réagit aux données d'entraînement (membres) et à celles qu'il n'a jamais vues (non-

membres), même si les jeux de données ne sont pas exactement les mêmes. Il est à noter que les algorithmes d'apprentissage automatique ont généralement une tendance à se comporter différemment en fonction de leur exposition antérieure à une donnée. Cette distinction, quoique délicate, peut être exploitée par un attaquant, même lorsqu'il dispose de peu d'informations ou travaille dans un environnement contraignant.

L'étude suggère aussi diverses sorties possibles du modèle cible pouvant servir à l'attaque : les seuls résultats prédits, les probabilités associées, ou encore les logits. Les auteurs ont évalué leur méthode sur une variété de jeux de données et de modèles (tels que les réseaux de neurones, les arbres de décision ou les SVM) et ont démontré que, même avec peu de données et des hypothèses limitées, l'attaque reste efficace. Cela souligne la vulnérabilité intrinsèque des modèles aux fuites d'information, même dans des cadres réalistes et contraints, comme le nuage ou les services *MLAAS* (Salem *et al.*, 2019).

Bu et collaborateurs ont proposé une nouvelle méthode d'attaque d'inférence d'appartenance qui ne nécessite pas de disposer de l'ensemble des informations génétiques d'un individu (Bu *et al.*, 2021). Grâce à cette méthode, l'adversaire peut s'appuyer sur des statistiques telles que les fréquences alléliques, c'est-à-dire la proportion d'un allèle donné dans une population, ou sur les réponses binaires de services Beacon, des interfaces publiques qui répondent par « oui » ou « non » à la question « Un allèle existe-t-il à une position spécifique dans une base de données génomique ? », afin de déterminer la présence d'un individu dans une base de données. Pour cela, ils exploitent les haplotypes, c'est-à-dire des combinaisons spécifiques d'allèles (variants génétiques) souvent transmises ensemble le long d'un même chromosome, ce qui offre plus de puissance statistique que l'analyse de variants pris isolément. Même si l'haplotype d'une personne n'est pas connu à l'avance, les auteurs montrent qu'il est possible de le reconstruire à partir de ces données résumées. Cette méthode soulève des inquiétudes majeures en matière de confidentialité, car elle montre que des informations sensibles peuvent être déduites même sans accès direct au génome complet.

Long et collaborateurs ont proposé une nouvelle perspective sur l'attaque d'inférence d'appartenance en se concentrant sur l'adversaire pragmatique qui cherche à maximiser l'utilité de l'attaque plutôt qu'à obtenir une couverture complète. Ils ont noté que, pour des modèles bien généralisés, les

membres et les non-membres d'un jeu de données sont souvent traités de manière similaire par le modèle, rendant les attaques classiques moins efficaces. Cependant, ils ont montré que l'adversaire pouvait cibler les données les plus sensibles, c'est-à-dire celles qui sont détectées plus facilement, afin de diminuer le nombre de non-membres prédits comme membres (faux positifs).

Pour identifier les enregistrements les plus vulnérables, ils ont recours à une méthode statistique fondée sur la valeur p, c'est-à-dire la probabilité d'observer un comportement au moins aussi extrême que celui mesuré sous l'hypothèse nulle (non-appartenance). Concrètement, il s'agit d'évaluer, pour une donnée cible, dans quelle mesure la réponse du modèle est compatible avec les comportements observés typiquement chez les membres ou chez les non-membres de l'ensemble d'entraînement. Une valeur p faible indique que la réaction du modèle à cette donnée ressemble beaucoup plus à celle qu'il aurait pour un exemple d'entraînement (membre) que pour une donnée inconnue (non-membre), ce qui laisse supposer que cette donnée a très probablement été utilisée lors de l'apprentissage du modèle. Long et collaborateurs ont également montré qu'il est possible d'atteindre une précision supérieure à 95 % dans certains sous-ensembles de données, même lorsque la précision globale d'une attaque semble faible (par exemple, autour de 50 %) (Long *et al.*, 2020).

Dans l'étude *Membership Inference Against DNA Methylation Database*, Hagestedt et collaborateurs examinent spécifiquement les attaques par inférence d'appartenance visant des bases de données de méthylation de l'ADN, un mécanisme épigénétique qui ajoute un groupe méthyle sur l'ADN afin de moduler l'activité des gènes sans en changer la séquence ; la méthylation joue ainsi un rôle important dans le développement, la régulation de l'expression génique et diverses pathologies (Schübeler, 2015). Les auteurs s'appuient largement sur les travaux existants concernant les attaques contre les données génomiques. Leurs résultats démontrent que ces attaques sont également efficaces contre les données de méthylation, en exploitant les statistiques résumées publiées. De plus, ils montrent que, même sans accès direct au profil de méthylation d'un individu, un attaquant disposant uniquement de ses variations génomiques peut inférer son appartenance à une base de données de méthylation. Cela est rendu possible grâce à la corrélation existante entre certains génotypes et les niveaux de méthylation observés à des positions spécifiques du génome. Cette approche met en lumière l'interconnexion croissante des risques pour la vie privée entre les données génomiques et épigénomiques (Hagestedt *et al.*, 2020).

Dans l'article *Differential Privacy Protection Against Membership Inference Attack on Machine Learning for Genomic Data*, les auteurs analysent le principal risque en matière de confidentialité posé par le partage de modèles formés à partir de données génomiques afin de prévoir un phénotype. Ils démontrent que les modèles d'apprentissage automatique sont vulnérables aux attaques par inférence d'appartenance, même lorsque seules les prédictions finales du modèle sont accessibles (cadre en boîte noire). Leur étude met en évidence que les adversaires peuvent exploiter les différences subtiles de comportement entre les membres et les non-membres de l'ensemble d'entraînement pour deviner si une donnée particulière a été utilisée pendant l'apprentissage. Ils ont aussi observé que certains facteurs, comme le surapprentissage du modèle et sa complexité, peuvent exacerber ces fuites d'information, ce qui augmente la probabilité de réussite de l'attaque (Chen *et al.*, 2020).

Comparativement, Shokri et collaborateurs ont proposé une méthode d'attaque plus efficace, mais également plus exigeante en termes d'information disponible pour l'adversaire. Leur approche repose sur la construction de plusieurs modèles d'ombre mimant la structure du modèle cible et entraînés sur des données issues de la même distribution. Ils ont ainsi démontré une efficacité remarquable, mais au prix d'hypothèses fortes sur la connaissance du modèle et des données (Shokri *et al.*, 2017).

Yeom et collaborateurs ont proposé une méthode très simple, peu coûteuse computationnellement, mais sensible à la régularisation et à la capacité du modèle à généraliser. Sa méthode est basée uniquement sur la perte, ce qui la rend très accessible, mais aussi très dépendante du degré de surapprentissage (Yeom *et al.*, 2018). L'approche de Salem et collaborateurs se situe entre les deux : elle relâche les hypothèses sur la distribution des données et la structure du modèle cible tout en conservant une performance comparable à celle de Shokri, avec une baisse de précision de quelques points de pourcentage. Elle met en évidence les risques réels pesant sur la vie privée et souligne le besoin de développer des modèles d'attaque plus robustes et adaptatifs. Elle montre surtout que même avec un seul modèle d'ombre et des données différentes, il est possible d'atteindre une performance proche, ce qui rend leur scénario plus réaliste pour des applications concrètes Salem *et al.* (2019).

2.4 Méthodes de défense contre les attaques MIA

En raison de l'augmentation du nombre d'attaques MIA, plusieurs stratégies de défense ont été proposées afin de limiter les fuites d'informations provenant des algorithmes d'apprentissage automatique. Ces méthodes diffèrent en termes de complexité, de niveau de confidentialité et d'impact sur les performances. Dans cette section, nous faisons un résumé des principales approches, en mettant en évidence leurs points forts et leurs limites. Nous précisons également pour quels scénarios d'attaque chaque méthode est la plus adaptée.

Plusieurs stratégies ont été proposées pour atténuer les attaques MIA. On peut les classer dans l'une ou l'autre des quatre grandes catégories suivantes : le masquage de la confiance, la régularisation, la confidentialité différentielle et la distillation des connaissances (Hu *et al.*, 2022).

2.4.1 Masquage de la confiance (*Confidence masking*)

Limiter ou altérer les informations divulguées par le modèle — par exemple en ne renvoyant que l'étiquette prédite, les k meilleures probabilités, ou en injectant du bruit dans le vecteur de sortie — réduit la surface informationnelle exploitable par un adversaire. Cette approche est particulièrement pertinente en boîte noire, notamment dans les services *MLAAS* où les attaquants exploitent les scores de confiance et les classements top-k. Elle se distingue par une mise en œuvre simple et un coût réduit ; de plus, la calibration des probabilités et le *label smoothing* atténuent la surconfiance, souvent ciblée par les attaques MIA. En revanche, son efficacité est moindre face aux attaques label-only et aux adversaires adaptatifs entraînés sur des sorties tronquées ou bruitées, et elle peut dégrader l'utilité des scores pour le seuillage, la supervision opérationnelle et l'explicabilité (Shokri *et al.*, 2017; Jia *et al.*, 2019; Li *et al.*, 2021; Choquette-Choo *et al.*, 2021). Deux notions sont particulièrement utiles dans ce contexte. La calibration des probabilités vise à ajuster les scores de sortie de manière à ce que, par exemple, une prédiction avec une confiance de 80 % soit correcte environ 8 fois sur 10 en pratique ; un modèle bien calibré est moins sujet à des surconfiances extrêmes sur les exemples d'entraînement. Le *label smoothing* consiste à remplacer les étiquettes one-hot $(1, 0, \ldots, 0)$ par des distributions légèrement adoucies (par exemple $(0{,}9, 0{,}1/(K-1), \ldots)$), ce qui empêche le modèle d'apprendre des frontières trop rigides et réduit l'écart de confiance entre membres et non-membres.

2.4.2 Régularisation

En réduisant la propension du modèle à mémoriser — par des pénalités L1/L2, du dropout, de l'augmentation de données ou de la régularisation adversariale explicitement ciblée sur l'appartenance — on réduit la différence de comportement entre les exemples connus (membres) et inconnus ; cette approche est généralement considérée comme la stratégie standard, y compris en *label-only*. Elle a l'avantage d'améliorer la généralisation sans changer l'API du modèle et avec une charge d'ingénierie raisonnable ; les versions adversariales peuvent encore diminuer la fuite d'information. Cependant, ses limites résident dans les régimes à haute dimension et faible effectif, où l'effet peut rester insuffisant ; par ailleurs, le réglage des hyperparamètres est complexe et des pénalisations/*dropout* trop intenses nuisent à la précision (Nasr *et al.*, 2018; Chang *et al.*, 2019; Salem *et al.*, 2019; Leino et Fredrikson, 2020).

2.4.3 Confidentialité différentielle

La confidentialité différentielle fournit un cadre formel pour garantir qu'un individu donné a un impact limité et contrôlé sur la sortie globale de l'algorithme. Plus précisément, un algorithme est (ϵ, δ)-différentiellement privé si, pour deux bases de données ne différant que par un individu, la distribution de ses sorties ne change que d'un facteur borné par ϵ (et δ) (Abadi *et al.*, 2016b). Ce paradigme est particulièrement indiqué lorsque le modèle est diffusé ou partagé, ou lorsqu'il est soumis à des contraintes réglementaires : il s'agit de la seule famille de défenses offrant des garanties formelles, avec un budget de confidentialité traçable.

Dans *DP-SGD*, on borne d'abord la norme des gradients individuels (*clipping*), puis on ajoute un bruit gaussien calibré avant l'agrégation, ce qui permet de suivre un budget de confidentialité (ϵ, δ) au cours de l'entraînement (Abadi *et al.*, 2016b). Dans *PATE* (*Private Aggregation of Teacher Ensembles*), plusieurs modèles enseignants sont entraînés sur des partitions disjointes des données ; leurs votes sur des exemples non étiquetés sont agrégés de manière bruitée pour entraîner un modèle élève, de sorte que l'influence de chaque individu reste limitée (Papernot *et al.*, 2018b).

En contrepartie, le compromis utilité–confidentialité est notable (des valeurs de ϵ faibles dégradent la performance), le réglage des hyperparamètres est complexe (*clipping*, échelle du bruit, taille de

lot, nombre d'itérations) et, en haute dimension, le niveau de bruit requis peut devenir important (Jayaraman et Evans, 2019; Shejwalkar et Houmansadr, 2021).

TABLE 2.2 – Comparaison des stratégies de défense contre les attaques MIA

Méthode	**Principe**	**Avantages**	**Limitations / Références**
Masquage de la confiance	Limitation des sorties (label seul, top-k, bruit sur les probabilités)	Facile à implémenter, efficace contre les attaques simples	Peut dégrader l'utilisabilité (Shokri *et al.*, 2017; Jia *et al.*, 2019)
Régularisation	Réduction du surapprentissage (L1/L2, dropout, data augmentation)	Renforce la généralisation, peu coûteux	Moins efficace en haute dimension (Nasr *et al.*, 2018; Chang *et al.*, 2019)
Confidentialité différentielle	Ajout de bruit pendant l'apprentissage (DP-SGD, PATE)	Garanties formelles de confidentialité	Perte de précision significative, tuning difficile (Abadi *et al.*, 2016b; Papernot *et al.*, 2018b)
Distillation	Transfert via un modèle enseignant	Réduction de la fuite sans bruit explicite	Complexité accrue, résultats variables (Shejwalkar et Houmansadr, 2021; Bernau *et al.*, 2021)

2.4.4 Distillation des connaissances

La distillation des connaissances consiste à entraîner un modèle élève sur les sorties de l'enseignant afin de lisser les signaux idiosyncratiques corrélés aux exemples d'entraînement et, ce faisant, de réduire l'écart de comportement entre membres et non-membres. Cette stratégie est pertinente lorsque la

restitution de probabilités de sortie est requise (contraintes produit/*MLAAS*) tout en évitant l'injection de bruit explicite. Sur le plan empirique, elle atténue la surconfiance et peut mieux préserver l'exactitude qu'une configuration DP stricte à ϵ faible, pour un niveau de résistance comparable. Ses limites tiennent à l'absence de garanties formelles, à une dépendance marquée au couple enseignant–élève, au schéma d'adoucissement des sorties et à la tâche, ainsi qu'à la possibilité pour un adversaire adaptatif de s'aligner si la géométrie des logits est insuffisamment modifiée (Shejwalkar et Houmansadr, 2021; Bernau *et al.*, 2021).

2.5 Conclusion

À travers ce chapitre, nous avons exploré les recherches existantes sur les attaques par inférence d'appartenance dans différents domaines de l'apprentissage automatique. Si ces attaques ont été largement étudiées dans des contextes comme la vision par ordinateur ou le traitement du langage, leur application aux données génomiques demeure encore marginale. Cette lacune ne reflète pas un manque d'intérêt, mais plutôt les défis spécifiques que posent les données génétiques : sensibles, complexes, difficilement partageables, riches en variables mais pauvres en échantillons. Par ailleurs, une majorité des travaux supposent un accès aux paramètres internes du modèle, ce qui est rarement réaliste dans des contextes biomédicaux. Ce constat met en lumière un espace encore peu exploré mais crucial : celui d'évaluer la vulnérabilité des modèles génomiques dans un cadre boîte noire, où seules les sorties du modèle sont accessibles à l'adversaire. Notre projet s'inscrit précisément dans cette perspective, en proposant une approche généralisable, reposant à la fois sur des modèles d'ombre entraînés à partir de phénotypes biologiquement corrélés au phénotype cible, et sur des modèles d'ombre construits à partir de jeux de données totalement indépendants selon une méthodologie de transfert de connaissances. Cette diversité dans la construction des modèles d'ombre permet d'anticiper les risques de divulgation, même dans des situations de contrôle d'accès restreint où l'attaquant ne possède que des données limitées sur le modèle cible. Ainsi, ce chapitre fournit le cadre conceptuel et l'état de l'art nécessaires pour analyser la vulnérabilité des modèles génomiques face aux attaques par inférence d'appartenance.

CHAPITRE 3
MÉTHODOLOGIE

Dans ce chapitre, la méthode utilisée pour mener une attaque par inférence d'appartenance en boîte noire sur des données génomiques sera décrite en détail. L'objectif principal est d'évaluer la vulnérabilité des modèles d'apprentissage automatique aux attaques de confidentialité en prédisant si un échantillon génomique donné faisait partie de l'ensemble de données d'entraînement du modèle. Pour y parvenir, la méthodologie est structurée en plusieurs étapes interconnectées.

Premièrement, on présente en détail l'ensemble de données utilisé, en explicitant les critères de sélection et de préparation adoptés. Nous détaillons ensuite le processus d'entraînement du modèle qui est la cible de l'attaque par inférence. De plus, nous expliquons comment créer et entraîner des modèles d'ombre, conçus pour imiter le comportement du modèle cible, ce qui est crucial pour générer des scénarios d'attaque réalistes. Par la suite, nous présentons et mettons en œuvre un modèle d'attaque par inférence d'appartenance robuste qui utilise les schémas détectés dans les résultats des modèles d'ombre pour déterminer l'état d'appartenance.

En outre, nous examinons minutieusement l'efficacité de l'attaque, en employant des critères bien définis pour mesurer sa précision et sa fiabilité. Dans tout ce chapitre, nous décrivons en détail l'environnement informatique et les outils spécifiques utilisés, ce qui facilitera la reproduction. Nous examinons également les considérations éthiques liées à la nature sensible des données génomiques, ainsi que les limites méthodologiques inhérentes à l'approche.

En définitive, le cadre méthodologique proposé fournit une approche structurée et rigoureuse pour analyser les menaces à la confidentialité posées par l'apprentissage automatique en génomique. Il permet d'éclairer les pratiques de protection des données et de préservation de la vie privée dans ce domaine.

3.1 Objectif de l'expérimentation

L'objectif principal de cette expérience est d'évaluer la faisabilité et la performance d'une attaque par inférence d'appartenance sur des données génomiques en boîte noire, en utilisant un modèle d'attaque généralisable. Afin d'assurer cette généralisabilité, l'attaque proposée est conçue pour être indépendante des individus cibles. Le modèle d'attaque ne repose donc pas sur la connaissance préalable des personnes potentiellement visées, mais apprend des schémas généraux issus de modèles d'ombre pour inférer l'appartenance. Dans le cadre de cette étude, la généralisabilité correspond à la capacité du modèle d'attaque à fonctionner de manière optimale sur les sorties d'un modèle cible inconnu, même si ce modèle a été entraîné sur un ensemble de données distinct ou qu'il utilise une architecture différente de celles utilisées lors de l'entraînement du modèle d'attaque. Elle mesure la résistance de l'attaque face à des modifications dans la nature des données, la structure du modèle ou la distribution des résultats. L'objectif est ainsi de pouvoir cibler n'importe quelle personne présentée au modèle cible, même si elle n'a jamais été vue auparavant, ni dans les données auxiliaires, ni lors de l'entraînement du modèle d'attaque. Il s'agit de prouver que, même en cas de ressources limitées et de restrictions concernant l'accès aux informations intermédiaires, il est possible de menacer la confidentialité des sujets représentés dans les échantillons d'apprentissage.

En ce qui concerne cette étude, plusieurs limites ont été rencontrées :

- la rareté des jeux de données génomiques publiques comportant un nombre d'échantillons suffisant ainsi que des annotations phénotypiques complètes.
- la nécessité d'un ensemble de données permettant de développer à la fois un modèle cible, un modèle d'ombre et un jeu d'attaque séparé.
- la grande quantité de caractéristiques, beaucoup plus élevée que le nombre d'échantillons, rend les modèles sensibles au surapprentissage.

Dans cette expérience, nous avons choisi de nous appuyer sur l'ensemble de données utilisé dans l'article de référence *"Differential Privacy Protection Against Membership Inference Attack on Machine Learning for Genomic Data"*(Chen *et al.*, 2020). Cet article constitue la base méthodologique principale de notre travail.

3.2 Présentation du jeu de données

Le jeu de données sélectionné est un ensemble de données de levures provenant de l'étude *"Genetic interactions contribute less than additive effects to quantitative trait variation in yeast"* (Bloom *et al.*, 2015). La levure (Saccharomyces cerevisiae) est un organisme modèle idéal pour ce type d'expérimentation. Elle possède un génome bien contrôlé, une faible complexité génétique et des phénotypes faciles à mesurer dans des conditions reproductibles. Elle est largement utilisée comme système modèle dans l'étude des mécanismes fondamentaux du vivant et des maladies humaines, grâce à la similitude qu'elle entretient avec les eucaryotes supérieurs et à sa maniabilité génétique (Dabas *et al.*, 2017; Poswal et Saini, 2017).

Le jeu de données contient des informations génétiques sur 4 390 individus issus d'un croisement entre deux souches de levure : une souche de laboratoire et une souche naturelle. Chaque individu a été génotypé sur plus de 28 820 marqueurs *SNP* et phénotypé pour une vingtaine de traits quantitatifs, majoritairement liés à la croissance cellulaire dans divers milieux.

Pour les besoins de notre expérimentation, nous nous sommes alignés sur le protocole de Chen *et al.* (2020) en choisissant comme phénotype cible le trait de croissance en présence de *sulfate de cuivre*. Parmi la vingtaine de traits quantitatifs disponibles, ce phénotype présente un bon compromis entre variabilité phénotypique et signal génétique, ce qui en fait un candidat adapté pour l'étude des attaques d'inférence d'appartenance. De plus, son utilisation nous permet de comparer plus directement nos résultats à ceux rapportés dans l'étude de référence de Chen *et al.* (2020).

Le phénotype, initialement mesuré comme une valeur quantitative de croissance, a été transformé en variable binaire en appliquant un seuil à la valeur de croissance. Ce seuil a été choisi de manière à obtenir deux classes de taille comparable (croissance « faible » vs « élevée »), de façon à limiter les déséquilibres de classes et à formuler la tâche comme un problème de classification supervisée binaire.

Sur les 4 390 individus initialement génotypés, nous avons d'abord exclu ceux ne disposant pas d'une mesure phénotypique valide pour le trait sulfate de cuivre. Ensuite, après application du seuillage et

sous-échantillonnage de la classe majoritaire pour équilibrer les étiquettes, le jeu de données final utilisé pour l'entraînement et l'évaluation du modèle cible contient 3 404 individus, dont 1 702 dans la classe positive (*résistants au sulfate de cuivre*) et 1 702 dans la classe négative (*sensible*).

3.2.1 Prétraitement des données

Les données de base se composent d'une matrice d'individus et de génotypes (utilisés comme caractéristiques), ainsi que d'un vecteur de phénotypes associés. La matrice de génotypes provient d'un fichier tabulaire, où chaque ligne correspond à un individu du panel de levure et chaque colonne représente un locus SNP. Les génotypes sont codés de manière binaire avec les entiers 1 et -1, correspondant respectivement aux allèles hérités des deux souches parentales utilisées dans le croisement expérimental. Ce mode de codage est souvent utilisé en analyse génétique, car il permet de représenter efficacement les effets cumulatifs. Dans un premier temps, les individus pour lesquels la valeur du phénotype est manquante sont retirés du jeu de données. Cette étape de filtrage permet de garantir que les matrices de génotypes et de phénotypes ont des dimensions compatibles et que chaque échantillon utilisé possède une annotation valide. Les phénotypes associés à ce jeu de données correspondent à la capacité de croissance des souches de levures dans divers milieux contenant des agents chimiques ou des sources de carbone spécifiques. Par exemple, on y trouve la réponse cellulaire en présence de sels métalliques (comme le chlorure de cobalt, le sulfate de cuivre ou le chlorure de magnésium), d'agents oxydants (comme le diamide), d'antibiotiques (comme la néomycine ou la zéocine) et même à la croissance sur différents substrats métaboliques (comme le lactose, le lactate ou le tréhalose). Pour éviter les biais causés par des classes déséquilibrées, un échantillonnage aléatoire de la classe majoritaire est fait. Cela permet d'obtenir un ensemble de données équilibré. Ensuite, les individus sont mélangés au hasard, mais avec une graine fixe pour assurer la reproductibilité. Les matrices génotypique et phénotypique finales peuvent alors être utilisées pour l'entraînement supervisé (Bloom *et al.*, 2015).

3.3 Aperçu du cadre expérimental

La figure 3.1 représente la méthodologie générale d'une attaque par inférence d'appartenance proposée par Shokri *et al.* (2017). Dans ce cadre, l'objectif de l'attaque par inférence d'appartenance

consiste à savoir si une donnée cible a été utilisée pour entraîner un modèle particulier ou non. L'adversaire ne connaît rien sur l'architecture du modèle cible ni sur ses paramètres (le cadre de boîte noire). Pour atteindre cet objectif, l'adversaire crée un modèle d'ombre destiné à imiter le comportement du modèle cible. Ce modèle est entraîné sur un jeu de données distinct de celui utilisé pour le modèle cible, mais censé provenir de la même distribution (ou d'une distribution similaire). Cette séparation stricte garantit que l'attaque repose uniquement sur la généralisation comportementale du modèle et non sur une fuite directe de données.

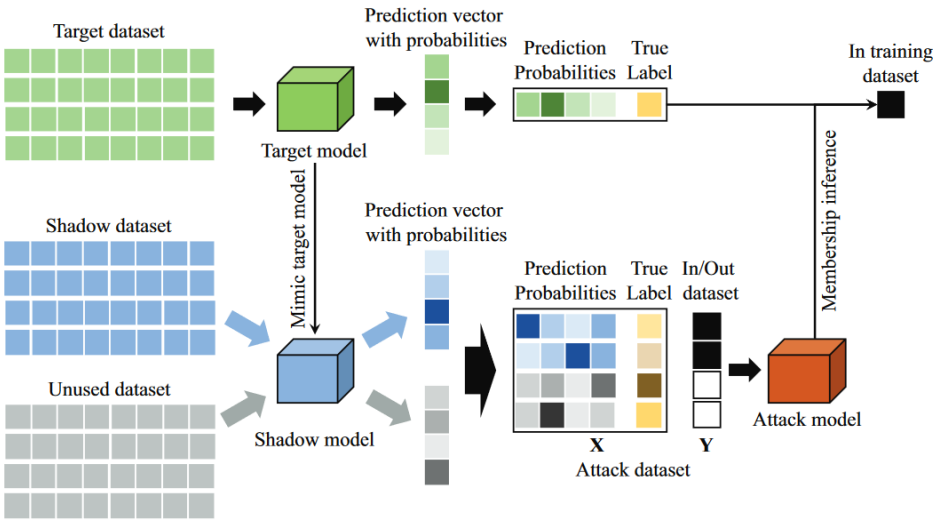

FIGURE 3.1 – Schéma général d'une attaque par inférence d'appartenance basée sur des modèles d'ombre (Chen *et al.*, 2020). La figure illustre les trois étapes principales : (1) entraînement du modèle cible sur ses données privées, (2) construction de modèles d'ombre sur des données auxiliaires co-distribuées, et (3) entraînement d'un modèle d'attaque à partir des sorties membres / non-membres. Elle sert de référence conceptuelle pour situer nos deux méthodologies par rapport au cadre proposé initialement par Shokri *et al.* (2017).

L'adversaire divise ce jeu de données auxiliaire en deux parties : l'une est utilisée pour l'entraînement du modèle de simulation des membres et l'autre pour le test, qui simule les non-membres. En interrogeant le modèle de simulation avec ces échantillons, il obtient les vecteurs de prédiction (c'est-à-dire les probabilités de classe) correspondant aux exemples connus comme étant des membres ou des non-

membres.

Ces sorties sont ensuite utilisées pour entraîner un modèle d'attaque, généralement un classificateur binaire, qui apprendra à différencier les résultats typiques d'un membre de ceux d'un non-membre. De cette façon, l'adversaire peut ensuite interroger le modèle cible avec un nouvel échantillon. En observant seulement les prédictions retournées, il peut ainsi déterminer si cet échantillon a probablement déjà fait partie de l'ensemble d'entraînement.

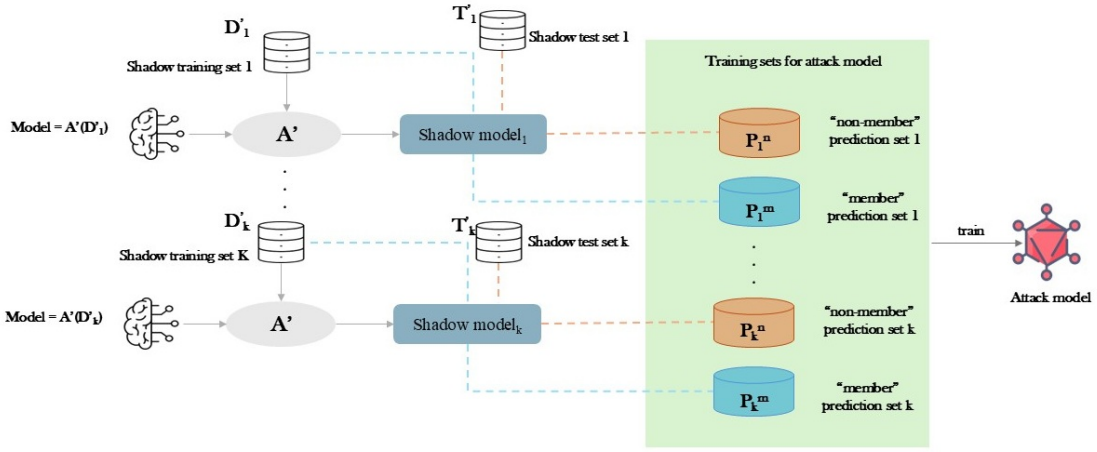

FIGURE 3.2 – Création de l'ensemble de données d'entraînement pour le modèle d'attaque à partir de plusieurs modèles d'ombre. Les sorties de ces modèles, évalués respectivement sur leurs jeux d'entraînement (membres) et de test (non-membres), sont agrégées pour former deux ensembles de vecteurs de probabilités P^m et P^n. Cette étape matérialise le lien entre le comportement de sur-confiance des modèles d'ombre et les étiquettes d'appartenance utilisées pour entraîner le classificateur d'attaque.

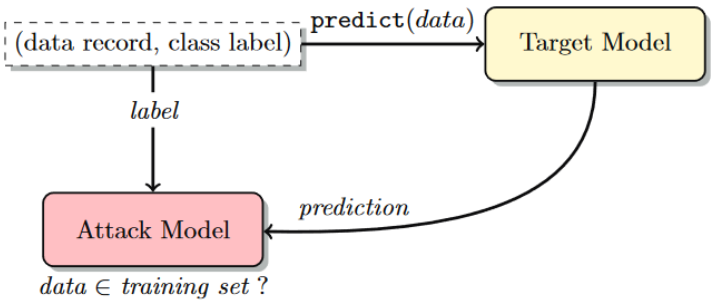

FIGURE 3.3 – Phase d'inférence de l'attaque par inférence d'appartenance (Shokri *et al.*, 2017). un exemple x est soumis au modèle cible et son vecteur de probabilités f(x) est fourni à un unique modèle d'attaque qui infère l'appartenance (membre / non-membre). L'étiquette vraie n'est pas utilisée comme entrée ; elle ne sert qu'à l'évaluation.

Une fois le modèle d'attaque entraîné, il peut être utilisé pour prédire si un échantillon donné a été vu par le modèle cible lors de l'entraînement. La figure 3.3 illustre cette phase d'inférence.

3.4 Modèle cible

Le modèle cible utilisé dans cette expérience est celui proposé par Chen *et al.* (2020) dans leur étude sur les attaques par inférence d'appartenance appliquée à des données génomiques en boîte blanche. Il s'agit d'un réseau de neurones convolutifs en une dimension (1D-*Convolutional neural network (CNN)*), une architecture particulièrement adaptée à la structure séquentielle des données génétiques. Ces dernières années, les réseaux de neurones convolutifs (CNN) ont gagné en popularité dans le domaine de la génomique, notamment pour la prédiction des phénotypes à partir des génotypes. Cette popularité est due à leur capacité à gérer des données extrêmement complexes, caractérisées par un grand nombre de variables (comme les SNPs) et un nombre limité d'échantillons. Dans ce type de configuration, les méthodes traditionnelles ont tendance à surapprendre, ce qui entrave la généralisation. À l'inverse, la structure hiérarchique des CNN, qui combine des couches de convolution, de pooling, de dropout et entièrement connectées, permet de limiter ce phénomène. Cette architecture facilite l'extraction automatique de motifs pertinents, tout en réduisant la dimension des données (Sehrawat *et al.*, 2023). Contrairement aux modèles linéaires traditionnels, les CNN peuvent captu-

rer des interactions complexes entre les variants génétiques, comme l'épistasie — des relations qui échappent souvent aux méthodes statistiques classiques (Guo et Li, 2023; Zhao *et al.*, 2016). Ils sont capables d'apprendre progressivement des représentations plus abstraites des données d'entrée, ce qui leur permet d'identifier des motifs génétiques récurrents associés à certains traits phénotypiques (Sehrawat *et al.*, 2023). En effet, les CNN se caractérisent non seulement par une précision accrue en matière de prévision, mais aussi par leur aptitude à produire des modèles plus stables et adaptables, même avec un ensemble restreint de données d'apprentissage. Ces avantages ont été démontrés dans plusieurs études appliquant les CNN à la prédiction des phénotypes à partir de différentes sources de variations génétiques (Gazestani et Lewis, 2019; Sehrawat *et al.*, 2023; Zhao *et al.*, 2016).

3.4.1 Description de l'architecture du modèle cible

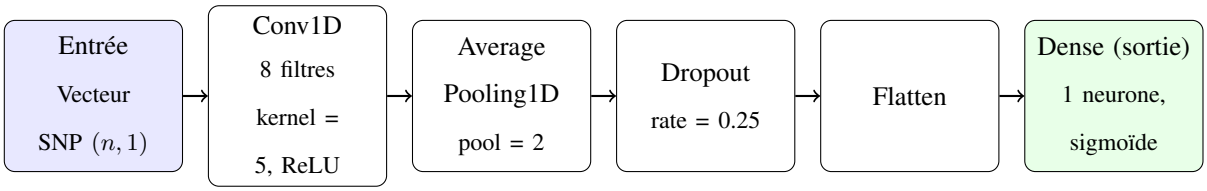

FIGURE 3.4 – Architecture détaillée du modèle cible (1D-CNN) utilisé pour la prédiction du phénotype sulfate de cuivre à partir des génotypes. La figure explicite la succession des couches (convolution, pooling, dropout, aplatissement, couche dense de sortie) et indique le type de tenseur manipulé à chaque étape.

Le modèle cible utilisé dans cette expérimentation est un réseau de neurones convolutifs unidimensionnels (1D-CNN), structuré de manière simple mais efficace pour la classification binaire à partir de données génomiques. Les réseaux CNN ont été introduits pour la première fois par LeCun *et al.* (1998) et sont depuis largement utilisés pour l'analyse de données structurées comme les images ou les séquences, en raison de leur capacité à extraire automatiquement des motifs locaux pertinents. Le modèle cible se compose des couches suivantes :

La première couche est une couche convolutionnelle 1D (Conv1D), qui applique un ensemble de filtres (ou noyaux) sur les séquences d'entrée afin de capturer les motifs locaux entre les loci SNPs. Cette couche utilise la fonction d'activation (*Rectified linear unit (RELU)*) pour introduire de la non-

linéarité, et une régularisation L1 est appliquée aux poids afin de limiter le surapprentissage. Ensuite, une couche de sous-échantillonnage moyenne (AveragePooling1D) permet de réduire la dimensionnalité en agrégeant les activations voisines, ce qui diminue la complexité du modèle et améliore la généralisation. Une couche de dropout est ensuite intégrée, désactivant aléatoirement une fraction des neurones lors de l'entraînement afin de renforcer la robustesse du modèle. Ensuite, la sortie est aplanie grâce à une couche Flatten, transformant la structure multidimensionnelle des activations précédentes en un vecteur unidimensionnel, qui peut être utilisé par la couche dense suivante. Finalement, la couche de sortie (Dense) contient un seul neurone dont l'activation dépend d'une fonction sigmoïde, produisant ainsi une probabilité d'appartenance à la classe positive. L'ensemble du réseau est entraîné à l'aide de l'optimiseur descente de gradient stochastique (*Stochastic gradient descent (SGD)*) et la fonction de perte est l'entropie croisée binaire, adaptée pour les tâches de classification binaire. Le modèle prend comme entrée un vecteur de génotypes de dimension $(n, 1)$, où n représente le nombre de SNPs. Chaque valeur représente le génotype d'un individu à une position spécifique. En sortie, le modèle génère une valeur scalaire comprise entre 0 et 1, représentant la probabilité qu'un échantillon appartienne à la classe positive du phénotype. Dans le cadre de cette expérimentation, le phénotype cible est la résistance au sulfate de cuivre, un trait binaire chez la levure.

Les hyperparamètres récapitulés dans le Tableau 3.1 reprennent ceux optimisés par Chen *et al.* (2020), ce qui nous permet de comparer directement nos résultats aux leurs.

Composant	Hyperparamètre	Valeur	Description
Conv1D	Nombre de filtres (`num_kernels`)	8	Nombre de noyaux appliqués sur la séquence d'entrée
	Taille du noyau (`kernel_size`)	5	Largeur du filtre utilisé
	Fonction d'activation	ReLU	Introduit la non-linéarité
	Régularisation	L1 ($\lambda = 0.001352$)	Encourage la parcimonie et réduit le surapprentissage
AveragePooling1D	Taille du pool	2	Réduction de la dimension par moyenne locale
Dropout	Taux de dropout (`dropout_rate`)	0.25	Fréquence de désactivation des neurones durant l'entraînement
Flatten	—	—	Aplatissement des activations pour la couche dense
Dense (Sortie)	Nombre de neurones	1	Sortie binaire (sigmoïde)
	Fonction d'activation	Sigmoïde	Renvoie une probabilité entre 0 et 1
Optimiseur	Type	SGD	Descente de gradient stochastique
	Taux d'apprentissage	0.01	Vitesse de mise à jour des poids
Entraînement	Nombre d'époques (`epochs`)	50	Nombre d'itérations sur le jeu d'entraînement
	Taille du batch (`batch_size`)	16	Nombre d'échantillons traités simultanément

TABLE 3.1 – Hyperparamètres et description du modèle cible (1D-CNN)

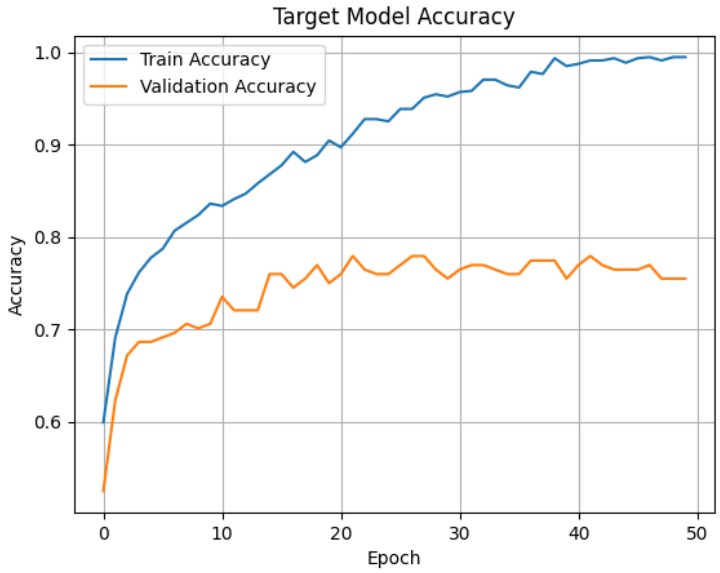

FIGURE 3.5 – Évolution de l'exactitude du modèle cible (1D-CNN) sur les jeux d'entraînement et de validation au cours des 50 époques d'apprentissage.

Afin d'évaluer la qualité de l'architecture choisie, nous avons analysé l'évolution de l'exactitude (*accuracy*) sur les jeux d'entraînement et de validation au cours des 50 époques d'apprentissage (voir Figure 3.5). L'exactitude d'entraînement augmente de manière régulière pour atteindre près de 0,99 à la fin de l'apprentissage, tandis que l'exactitude de validation progresse plus modérément et se stabilise autour de 0,75–0,78 après une trentaine d'époques. L'écart observé entre les deux courbes reflète un surapprentissage léger, attendu compte tenu du faible nombre d'échantillons et du grand nombre de SNPs, mais ne s'accompagne d'aucune dégradation soudainement marquée des performances de validation. Cela montre que le modèle conserve une capacité de généralisation satisfaisante sur les données de levure.

Dans les expériences qui suivent, nous conservons les poids correspondant à l'époque présentant la meilleure exactitude de validation, afin de limiter l'effet de surapprentissage. Ces observations confirment que l'architecture 1D-CNN retenue est suffisamment expressive pour capturer les signaux génétiques liés au phénotype sulfate de cuivre, tout en procurant une performance de généralisation adéquate pour les besoins de l'étude.

3.5 Synthèse comparative

Ce mémoire présente deux stratégies d'attaque distinctes pour l'inférence d'appartenance. La première repose sur la construction de modèles d'ombre à partir de phénotypes corrélés, tandis que la seconde utilise des ensembles de données hétérogènes et indépendants. Ces deux approches offrent des perspectives complémentaires pour l'inférence d'appartenance. Dans les sections suivantes, nous décrivons en détail la méthodologie et le protocole expérimental associés à chacune de ces deux approches.

TABLE 3.2 – Comparaison des deux méthodologies d'attaque

Critère	**Méthode 1 : Modèles d'ombre corrélés**	**Méthode 2 : Attaque généralisée**
Source des modèles d'ombre	Données du même domaine, phénotypes corrélés	Données externes hétérogènes (images, textes, etc.)
Hypothèse principale	Corrélation génétique suffisante entre phénotype cible et auxiliaires	Existence de motifs génériques dans les sorties entre membres et non-membres
Avantage principal	Proximité biologique, meilleure représentativité du modèle cible	Indépendance vis-à-vis du domaine, réutilisation possible sur plusieurs cibles
Limite principale	Nécessite une bonne sélection des phénotypes auxiliaires	Écart de distribution entre les données d'entraînement et celles du modèle cible
Robustesse	Forte pour des phénotypes bien choisis, mais limitée hors corrélation	Surprenante malgré l'hétérogénéité des données d'entraînement
Complexité computationnelle	Moyenne (entraîner quelques modèles similaires)	Élevée (multiples jeux de données et post-traitement)
Applicabilité aux services boîte noire	Possible si des phénotypes auxiliaires sont accessibles	Applicable si les sorties probabilistes du modèle sont disponibles

3.6 Stratégies d'entraînement du modèle d'ombre

L'objectif principal de cette section est d'explorer différentes approches pour générer les données nécessaires à l'entraînement du modèle d'attaque, dans un contexte réaliste de boîte noire où l'adversaire n'a accès qu'aux sorties du modèle cible. Nous avons mis en œuvre et comparé deux stratégies distinctes :

- Attaque par modèle d'ombre : cette méthode consiste à créer un modèle d'ombre spécifique pour imiter le comportement du modèle cible. Pour ce faire, nous avons combiné plusieurs jeux de données, tous basés sur les mêmes profils génétiques, mais associés à des phénotypes différents. Cette diversité phénotypique permet d'entraîner un modèle d'ombre capable de généraliser le comportement du modèle cible.
- Attaque par transfert de connaissances généralisée : contrairement à la première méthode, celle-ci ne cherche pas à reconstruire la structure ou le comportement du modèle cible. Elle se base uniquement sur les prédictions du modèle cible sur des données auxiliaires pour générer directement les étiquettes de membres et non-membres. Par conséquent, elle évite toute modélisation intermédiaire.

Ces deux stratégies ont été comparées afin d'évaluer leur efficacité respective pour entraîner un modèle d'attaque capable de généraliser et d'identifier correctement l'appartenance des échantillons, même avec une connaissance limitée du modèle cible.

3.6.1 Méthode 1 : méthodologie d'attaque par modèle d'ombre

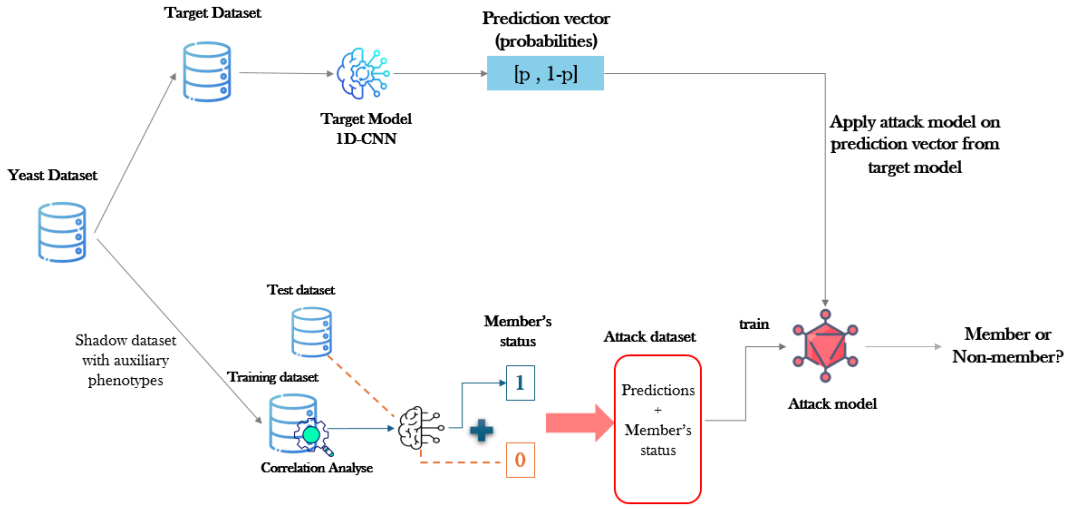

FIGURE 3.6 – Pipeline de la Méthode 1 : attaque généralisée par modèle d'ombre basé sur des phénotypes auxiliaires corrélés. La figure montre comment les données génomiques et les phénotypes auxiliaires sont utilisées pour entraîner un modèle d'ombre, produire des sorties membres / non-membres, puis former un modèle d'attaque appliqué ensuite aux prédictions du modèle cible. Elle illustre le rôle central de la corrélation phénotypique pour rapprocher le comportement du modèle d'ombre de celui du modèle cible.

Cette première méthode consiste à concevoir un modèle d'ombre généralisé visant à simuler le fonctionnement d'un modèle cible entraîné sur un phénotype spécifique. Dans notre exemple, le modèle cible est un réseau de neurones convolutifs unidimensionnels (1D-CNN) entraîné pour prédire la présence ou l'absence d'un phénotype particulier (sulfate de cuivre) à partir de données de génotypes.

Étant donné la taille limitée du jeu de données génétiques disponibles, nous avons choisi de rester dans la même répartition génétique, mais en utilisant différents phénotypes pour entraîner notre modèle. Pour maximiser cette sélection, nous avons créé une matrice de corrélation entre les 19 autres phénotypes mesurés sur le même groupe d'individus, en excluant le phénotype cible (sulfate de cuivre). Cette étude visait à déterminer le phénotype le plus similaire statistiquement aux autres, afin de permettre au modèle d'ombre de mieux approximer le comportement du modèle cible à partir d'un phé-

notype biologiquement et structurellement proche.

Bien qu'elle soit basée sur une analyse statistique préalable, il faut noter que le phénotype cible n'a pas été utilisé pour construire le modèle d'ombre ni pour calculer les corrélations. Ainsi, l'adversaire n'a aucune information spécifique sur les étiquettes du modèle cible, mais seulement un accès aux données de même distribution associées à des phénotypes différents. D'un point de vue méthodologique, nous supposons donc que les phénotypes auxiliaires sont mesurés sur la même cohorte que le phénotype cible, mais dans des contextes expérimentaux différents (autres milieux de culture, autres stress, etc.). Ce choix ne constitue pas une fuite artificielle d'information au profit de l'attaquant, dans la mesure où les étiquettes du phénotype cible ne sont jamais réutilisées pour entraîner les modèles d'ombre. Au contraire, il reflète un scénario réaliste où un même individu peut apparaître dans plusieurs études ou essais cliniques, avec des phénotypes multiples mesurés sur un même génome. Dans ce cadre, l'utilisation de ces phénotypes auxiliaires permet à l'attaquant de tirer parti de signaux génétiques partagés entre traits corrélés, tout en respectant le cadre boîte noire : seules des données externes ou auxiliaires, distinctes des étiquettes du modèle cible, sont exploitées pour construire les modèles d'ombre. Il est important de noter que, bien que les génotypes des modèles d'ombre et du modèle cible proviennent de la même cohorte, les ensembles utilisés pour évaluer l'attaque sont strictement séparés de ceux utilisés pour l'entraînement. En particulier, les exemples «membres» et «non-membres» du modèle cible sont construits à partir de sous-ensembles distincts, et les modèles d'ombre ne sont jamais entraînés avec les étiquettes du phénotype cible. Cette séparation vise précisément à limiter le surapprentissage sur un jeu de données particulier et à préserver la capacité de généralisation de l'attaque dans le cadre expérimental étudié.

Ce cadre est inspiré d'usages réels en médecine personnalisée, où des modèles prédisent la réponse à un traitement à partir du génotype. Un adversaire peut connaître la nature de la tâche (p. ex. « réponse au médicament X ») et disposer de données publiques/auxiliaires sur la même population, associées à d'autres phénotypes (p. ex. réponses à traitements voisins), sans recourir aux exemples d'entraînement du modèle cible. L'exploitation de phénotypes corrélés permet alors d'entraîner un modèle d'ombre généralisable, qui capture des signaux de surconfiance ou de mémorisation similaires à ceux du modèle cible, tout en respectant les contraintes de réalisme (boîte noire) et de confidentialité.

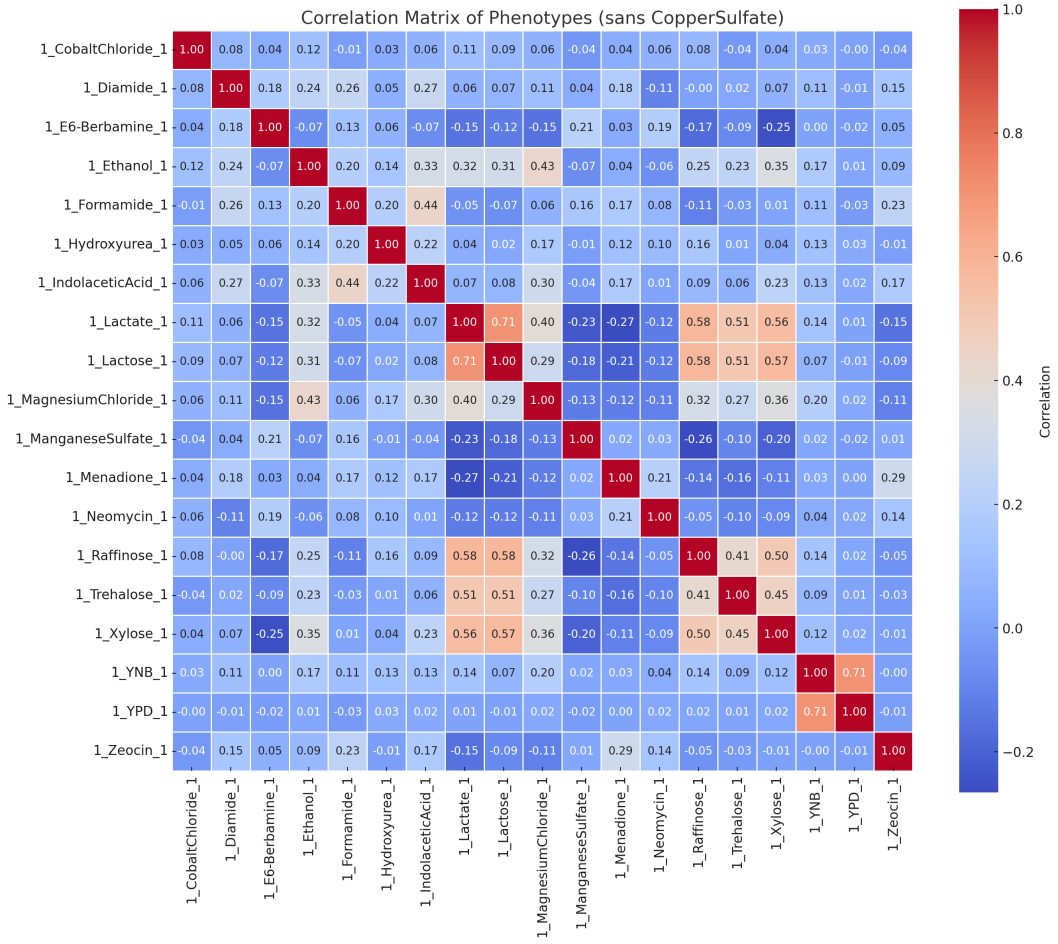

FIGURE 3.7 – Matrice de corrélation de Pearson entre 19 phénotypes mesurés sur un même ensemble de génotypes de levure. Chaque case représente la corrélation entre une paire de traits, les valeurs élevées (en valeur absolue) signalant des phénotypes susceptibles de partager des déterminants génétiques communs. Cette visualisation sert de base à la sélection de phénotype auxiliaire utilisé pour entraîner le modèle d'ombre dans la Méthode 1.

Cette matrice (voir Figure 3.7) a été construite à partir des valeurs phénotypiques mesurées pour chaque individu, en calculant les coefficients de corrélation de Pearson entre chaque paire de phénotypes. L'objectif est d'identifier les relations statistiques existantes entre différentes conditions environnementales ou stress biologiques, afin de sélectionner des phénotypes auxiliaires pertinents pour la construction d'un modèle d'ombre. Une valeur absolue élevée indique une variation phénotypique

potentiellement similaire, ce qui peut améliorer la capacité du modèle d'ombre à approximer le comportement du modèle cible.

Après avoir construit la matrice, nous avons déterminé les phénotypes les plus dominants en additionnant les valeurs absolues des coefficients de corrélation pour chaque phénotype (ligne/colonne) dans cette dernière. Les phénotypes Lactate, Lactose, Xylose, Raffinose et Magnesium Chloride se sont distingués par les sommes les plus élevées (respectivement 5,48, 5,10, 4,98 , 4,93 et 4,60), indiquant une forte similarité statistique avec l'ensemble des autres traits. Parmi ces phénotypes fortement corrélés, nous avons finalement retenu Xylose comme phénotype auxiliaire principal pour l'implémentation expérimentale. Ce choix est motivé à la fois par sa corrélation élevée avec l'ensemble des autres traits et par des contraintes computationnelles, qui nous ont conduit à concentrer l'analyse détaillée sur un seul phénotype représentatif. Cette sélection permet d'entraîner le modèle d'ombre sur des phénotypes situés statistiquement au centre du réseau de corrélations, ce qui favorise l'apprentissage de patrons de variation phénotypique plus généraux et représentatifs. Le phénotype auxiliaire utilisé dans cette méthode ont donc été choisis selon une approche statistique rigoureuse. En effet, la matrice de corrélation a permis d'identifier ceux présentant les similarités les plus fortes avec l'ensemble des traits mesurés. Le principe implicite est que les phénotypes biologiquement proches ont des schémas de variation génétique similaires. Cela rend les prédictions de modèle d'ombre formé sur ces caractéristiques plus appropriées pour simuler le comportement du modèle cible.

Le modèle d'ombre choisi dans cette méthodologie est un classificateur de type *régression logistique*, entraîné à partir des génotypes associés au phénotype auxiliaire sélectionné. La régression logistique est couramment utilisée en génomique pour modéliser la relation entre des variants génétiques (tels que les SNPs) et des phénotypes binaires (comme la présence ou l'absence d'une maladie). Elle permet de mettre en évidence les variants significativement associés à un trait et de fournir des coefficients interprétables, qui représentent l'impact de chaque variant sur la probabilité d'occurrence du phénotype (Wu *et al.*, 2009; Sperandei, 2014).

Chaque modèle d'ombre est entraîné à l'aide d'exemples connus (provenant du jeu d'entraînement) et évalué à l'aide d'exemples inconnus (provenant du jeu de test), ce qui donne deux ensembles de sorties : l'un pour les membres et l'autre pour les non-membres. Dans notre protocole, le modèle

d'ombre est entraîné sur des sous-ensembles dédiés, distincts de ceux utilisés pour l'entraînement et l'évaluation du modèle cible. Les exemples issus de l'ensemble d'entraînement des modèles d'ombre sont considérés comme « membres », tandis que ceux issus de leur ensemble de test jouent le rôle de « non-membres » pour la construction du jeu de données d'attaque.

Après avoir entraîné des modèles d'ombre avec les données de levure, ils ont été utilisés pour générer les prédictions nécessaires à la construction du jeu de données d'attaque. Chaque modèle d'ombre a fourni des estimations de probabilité, tant pour les échantillons d'entraînement (considérés comme membres) que pour les échantillons de test (non-membres). Ces prédictions reflètent la manière dont le modèle traite les données qu'il a vues par rapport à celles qu'il n'a jamais rencontrées, une distinction essentielle pour une attaque d'inférence d'appartenance.

TABLE 3.3 – Conditions d'entraînement du modèle d'ombre (Méthode 1)

Paramètre	**Valeur**
Type de modèle	Régression logistique
Pénalité	L2
C (inverse de la régularisation)	1.0
Solveur	lbfgs
Nombre maximal d'itérations	1000

Chaque exemple est alors représenté par son vecteur de sortie (par exemple, la probabilité associée à la classe positive pour un problème binaire). Il porte également une étiquette : 1 s'il a été vu par le modèle (membre), 0 sinon (non-membre). L'ensemble de ces vecteurs (environ 1 000 exemples équilibrés) constitue le jeu d'entraînement du modèle d'attaque.

Les vecteurs issus de l'ensemble des modèles d'ombre sont ensuite concaténés pour former le jeu de données d'entraînement du modèle d'attaque. Ce modèle est entraîné pour apprendre à distinguer les sorties typiques d'un échantillon membre de celles d'un non-membre. Nous avons utilisé ici un classi-

fieur forêt aléatoire (Breiman, 2001), connu pour sa robustesse, sa capacité à capturer des interactions complexes et sa bonne performance sur des jeux de données de taille moyenne. Une validation croisée n'a pas été jugée nécessaire dans ce contexte, car le jeu de données est équilibré, le modèle n'est pas fortement paramétré, et les performances se stabilisent rapidement lors de l'entraînement.

De plus, le choix d'un modèle d'ombre plus simple que le modèle cible (sous forme de régression logistique) est volontaire : il reflète un attaquant réaliste qui ne cherche pas à reproduire fidèlement l'architecture interne du 1D-CNN, mais à capturer des tendances générales entre membres et non-membres. De la même manière, l'utilisation d'une forêt aléatoire comme modèle d'attaque relève d'une approche pragmatique dans un scénario de boîte noire : l'adversaire ne connaît ni la structure exacte du modèle cible ni ses hyperparamètres, et s'appuie donc sur un classificateur robuste, stable et peu paramétré pour capturer les schémas discriminants dans les sorties des modèles d'ombre.

TABLE 3.4 – Conditions d'entraînement du modèle d'attaque

Paramètre	**Valeur**
Type de modèle	Forêt aléatoire
Profondeur maximale	2
Random state	42
Taille du jeu d'attaque	$\sim$1 000 exemples (membres et non-membres)
Métriques d'évaluation	Accuracy, Precision, Recall, F1, ROC AUC

3.6.2 Méthode 2 : méthodologie d'attaque par transfert de connaissances généralisée

Dans le cadre de ce travail, l'approche par transfert de connaissances généralisées a été retenue pour plusieurs raisons pratiques et méthodologiques.

Tout d'abord, la quantité de données génomiques de levure disponible était limitée. Pour mener une attaque par inférence d'appartenance, il est nécessaire de diviser ces données en trois ensembles distincts : l'un pour entraîner le modèle cible, un autre pour créer les modèles d'ombre et le dernier pour tester l'attaque. Cela aurait considérablement réduit la quantité d'échantillons utilisables pour chaque étape. Ensuite, puisque notre scénario est réaliste et du genre boîte noire (l'attaquant n'a accès qu'aux

sorties du modèle cible), et que l'objectif est de concevoir un modèle d'attaque généralisé, la méthodologie présentée dans l'article *ML-Leaks : Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models* (Salem *et al.*, 2019) s'est avérée très appropriée. Cette méthode permet effectivement de résoudre la contrainte liée au nombre limité d'échantillons disponibles, en s'appuyant sur des ensembles de données externes pour entraîner les modèles d'ombre, sans nécessiter un accès direct aux données d'entraînement du modèle cible.

Cette méthode fonctionne de manière conceptuellement distincte des attaques par modèles d'ombre traditionnelles, que nous allons détailler ci-dessous. Dans cette méthodologie, l'attaquant adopte une approche différente de la méthode traditionnelle des attaques par modèles d'ombre. Contrairement à l'approche initiale de Shokri *et al.* (2017), qui vise elle aussi l'inférence d'appartenance mais apprend l'attaquant via des modèles d'ombre "miroirs" entraînés sur des données co-distribuées et étiquetées, notre approche suit Salem *et al.* (2019) et entraîne des modèles d'ombre sur des ensembles externes potentiellement hétérogènes, puis extrait des résumés statistiques des sorties (p. ex. top-k, entropie, pertes) pour entraîner le classifieur d'attaque sans accès aux données d'entraînement ni à la structure interne du modèle cible. Comme l'attaquant ne possède ni les données d'entraînement du modèle cible ni sa structure exacte, il utilise donc un ou plusieurs ensembles de données externes, qui peuvent être tirés de sources très diverses, par exemple des images, du texte ou encore des transactions. Alors, chaque modèle d'ombre est entraîné sur un sous-ensemble de ses données (considéré comme les "membres") et il est ensuite évalué à partir d'une autre partie (considérée comme les "non-membres").

Les ensembles de données utilisés dans cette méthodologie ont été sélectionnés selon plusieurs critères : leur accessibilité publique, leur diversité structurelle (images, texte, données tabulaires, etc.), ainsi que leur usage fréquent dans les études sur les attaques MIA, notamment dans le cadre de l'approche de Salem *et al.* (2019). Ces jeux permettent de simuler différents types de comportements modèles sans dépendre des données cibles.

Parmi ces ensembles, on retrouve notamment :

— Purchase-100 [1] : un jeu de données transactionnel composé de profils d'achats binaires répartis

1. https ://www.kaggle.com/c/acquire-valued-shoppers-challenge/data

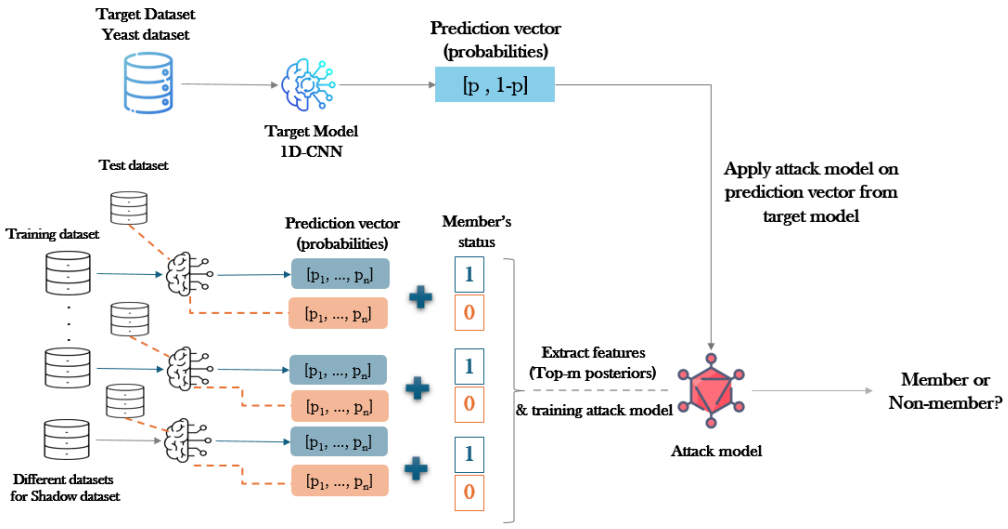

FIGURE 3.8 – Pipeline de la Méthode 2 : attaque par transfert de connaissances généralisée. Des modèles d'ombre sont d'abord entraînés sur plusieurs jeux de données externes hétérogènes (images, texte, données tabulaires, etc.) afin de capturer des motifs génériques de différences entre membres et non-membres dans l'espace des postériors. Les caractéristiques dérivées (top-3 postériors) servent ensuite à entraîner un modèle d'attaque, qui est finalement appliqué aux sorties du modèle cible de levure, sans jamais avoir vu ses données d'entraînement.

en 100 classes. Il est notamment utilisé dans les expériences de Salem *et al.* (2019).

— Adult [2] : un jeu de données tabulaires couramment utilisé dans l'inférence d'appartenance, provenant du recensement des États-Unis. Il contient une catégorisation binaire.

— MNIST [3] : un jeu d'images manuscrites en niveaux de gris de chiffres (0 à 9), contenant 10 classes.

— CIFAR-10 et CIFAR-100 [4] : deux jeux de données d'images couleur, contenant respectivement 10 et 100 classes, couvrant des objets variés (animaux, véhicules, etc.).

— Location [5] : un ensemble de données de géolocalisation extrait des points de check-in Fours-

2. https ://archive.ics.uci.edu/dataset/2/adult
3. https ://www.kaggle.com/datasets/hojjatk/mnist-dataset
4. https ://www.cs.toronto.edu/ kriz/cifar.html
5. https ://sites.google.com/site/yangdingqi/home/foursquare-dataset

quare, utilisé dans plusieurs études récentes sur la confidentialité des modèles (Yang *et al.*, 2014).

— News [6] : un ensemble de données textuelles composé de 20 forums de discussion différents, utile pour l'entraînement de modèles à partir de séquences de texte.

Du point de vue de la génomique, ces ensembles externes ne fournissent aucune information directe sur la levure. Leur fonction consiste plutôt à offrir un éventail de tâches de classification supervisée dans lesquelles on observe le même phénomène structurel que dans les attaques MIA : les exemples vus pendant l'entraînement ont tendance à produire des résultats plus confiants (postérieurs plus concentrés, entropie plus faible) que les exemples inédits. L'hypothèse de transfert est donc que ces motifs de sur-confiance dans l'espace des probabilités de sortie sont en grande partie indépendants du domaine et peuvent être appris à partir de données arbitraires, puis réutilisés avec le modèle cible de levure. Par conséquent, l'utilité de ces jeux d'images, de texte ou de transactions ne réside pas dans leur signification sémantique, mais dans leur capacité à saisir des schémas génériques de comportement de membre ou de non-membre d'un modèle d'apprentissage.

TABLE 3.5 – Modèles d'ombre utilisés pour chaque ensemble de données

Nom du dataset	**Type de données**	**Modèle d'ombre utilisé**
MNIST	Images (10 classes)	CNN (2 Conv2D + MaxPooling + Dense)
CIFAR-10	Images (10 classes)	CNN (blocs Conv2D avec BatchNorm et Dropout)
CIFAR-100	Images (100 classes)	EfficientNetB0 préentraîné (ImageNet)
Adult	Données tabulaires	Fôret aléatoire (Scikit-learn)
Purchase-100	Données tabulaires	MLP (2 couches Dense de 1024 neurones + Softmax)
20 Newsgroups	Texte	Multinomial Naive Bayes avec TF-IDF
Location (TIST)	Coordonnées géographiques	MLP (3 couches avec Dense, Dropout et Softmax)

6. https ://www.kaggle.com/datasets/everydaycodings/global-news-dataset

TABLE 3.6 – Caractéristiques des ensembles de données externes utilisés (Méthode 2)

Nom du dataset	**Type de données**	**Nb d'attributs (ordre de grandeur)**	**Nb de classes**
MNIST	Images 28×28 en niveaux de gris	784 pixels	10 chiffres (0–9)
CIFAR-10	Images couleur $32 \times 32 \times 3$	$\sim 3\,000$ pixels	10 catégories d'objets
CIFAR-100	Images couleur $32 \times 32 \times 3$	$\sim 3\,000$ pixels	100 catégories d'objets
Adult	Données tabulaires (recensement)	$\sim$ 100 attributs après encodage	2 classes de revenu
Purchase-100	Vecteurs binaires de comportement d'achat	~ 600 produits	100 segments de clients
20 Newsgroups	Texte (représentation TF-IDF)	$\sim 5\,000$ termes fréquents	20 groupes de discussion
Location (TIST)	Séquences de check-ins géolocalisés	Quelques dizaines de descripteurs dérivés	Plusieurs dizaines de lieux

Pour s'adapter aux particularités de chaque ensemble de données, des algorithmes d'apprentissage adaptés ont été choisis pour entraîner les modèles d'ombre. Les modèles sont présentés dans le tableau 3.5. Pour sélectionner les modèles d'ombre appropriés pour chaque ensemble de données, plusieurs facteurs ont été pris en compte : (1) la nature des données (images, tabulaires, texte, etc.), (2) la complexité du problème (nombre de classes, linéarité des relations), et (3) la littérature existante sur les performances des modèles d'inférence d'appartenance. Les CNN ont été privilégiés pour les données de type image, car ils permettent une extraction hiérarchique des motifs visuels (Shokri *et al.*, 2017; Salem *et al.*, 2019; Nasr *et al.*, 2018). Les *Multilayer Perceptron (MLP)* sont adaptés aux données structurées à haute dimension comme Purchase-100 (Salem *et al.*, 2019). Enfin, les forêts

aléatoires et les modèles bayésiens sont efficaces pour les jeux de données plus simples ou de plus petite taille (Yeom *et al.*, 2018; Jayaraman et Evans, 2019). Cette approche ciblée permet d'optimiser les performances de chaque modèle d'ombre tout en garantissant la diversité des comportements utilisés pour entraîner le modèle d'attaque.

Contrairement aux approches traditionnelles qui construisent plusieurs modèles d'ombre pour simuler le comportement du modèle cible, notre méthodologie adopte une version simplifiée, dans laquelle un seul modèle d'ombre est créé pour chaque ensemble de données. Cette simplification permet non seulement de réduire la complexité computationnelle, mais aussi de maîtriser la variabilité entre les jeux de données hétérogènes. Elle assure également une couverture suffisante des différences entre membres et non-membres.

Une fois les prédictions des modèles d'ombre collectées (sous forme de postérieurs), un modèle d'attaque est entraîné pour distinguer les échantillons membres des non-membres. Dans notre implémentation, chaque exemple est représenté par un vecteur de caractéristiques statistiques extraites des sorties du modèle, plus précisément les trois probabilités les plus élevées parmi toutes les classes prédites (aussi appelées top-3 postérieurs). Cette méthode d'extraction permet de résumer l'information de sortie tout en réduisant la dimensionnalité, ce qui améliore la stabilité du modèle d'attaque. Pour constituer le jeu de données d'entraînement de l'attaque, les sorties des modèles d'ombre sont étiquetées en fonction de leur appartenance : les échantillons vus lors de l'entraînement du modèle d'ombre sont étiquetés comme membres (1), et ceux utilisés pour le test comme non-membres (0). Ces vecteurs sont ensuite concaténés pour entraîner un classificateur forêt aléatoire, choisi pour sa capacité à capturer des interactions non linéaires, sa robustesse aux variables redondantes et sa bonne généralisation sur des jeux de taille moyenne (Breiman, 2001).

Dans notre cas, le jeu d'entraînement du modèle d'attaque est construit à partir de sept jeux de données hétérogènes (Purchase, Adult, CIFAR-10/100, MNIST, Location, News). Chaque dataset a fourni environ 2000 à 10000 échantillons membres et non-membres, soit un total combiné de plus de 40 000 exemples d'entraînement. Le modèle d'attaque est validé sur des exemples générés à partir du modèle cible réel, à savoir le CNN entraîné sur les données génomiques de levure. Pour cela, un ensemble de membres (issus de l'entraînement du modèle cible) et un ensemble de non-membres (issus d'un jeu

de données totalement disjoint) sont soumis au modèle cible, et leurs prédictions sont transformées en top-3 postériors comme décrit ci-dessus.

Ce protocole d'évaluation permet de mesurer la capacité de généralisation de l'attaque, c'est-à-dire sa capacité à détecter l'appartenance d'un échantillon sans jamais avoir vu le modèle cible ou ses données d'origine. L'ensemble final de test est équilibré (50% membres / 50% non-membreset comprend environ 2 000 échantillons.

3.7 Protocole de validation

Dans cette section, nous décrivons les stratégies de validation utilisées pour chacune des deux méthodologies proposées.

3.7.1 Méthodologie 1 : Modèles d'ombre basés sur des phénotypes corrélés

- Le jeu de données initial contient 4 390 individus génotypés. Après filtrage des échantillons ne disposant pas d'une mesure phénotypique valide pour le phénotype cible (sulfate de cuivre), il reste 3 404 individus.
- Après binarisation du phénotype et équilibrage des classes, le jeu de données final comporte 3 404 échantillons, soit 1 702 individus dans la classe positive et 1 702 dans la classe négative.
- Environ 70 % de ces individus (2 383 échantillons) sont utilisés pour entraîner et évaluer le modèle cible. Ce sous-ensemble est ensuite divisé en 80 % pour l'entraînement (1 907 échantillons) et 20 % pour le test du modèle cible (476 échantillons).
- Les 30 % restants (1 021 échantillons) sont conservés comme données non vues par le modèle cible. Ces données servent de base pour constituer les exemples « non-membres » dans l'évaluation de l'attaque.
- Pour la construction du modèle d'attaque, un ensemble équilibré est formé à partir de :
  - exemples « membres » : échantillons issus de l'ensemble d'entraînement du modèle cible ;
  - exemples « non-membres » : échantillons tirés au hasard parmi les 1 021 profils réservés et jamais vus par le modèle cible.
- Dans cette première méthodologie, les modèles d'ombre sont entraînés sur des phénotypes

auxiliaires mesurés sur la même cohorte de levures, mais en utilisant des sous-ensembles strictement séparés de ceux employés pour l'entraînement et l'évaluation du modèle cible, afin d'éviter toute fuite directe d'information et de mesurer la généralisation réelle de l'attaque.

3.7.2 Méthodologie 2 : Transfert de connaissances généralisé

Le jeu de données génomiques initial contient 4 390 individus. Après filtrage des phénotypes manquants pour le trait *sulfate de cuivre*, il reste 3 404 individus utilisables, qui sont répartis comme suit :

- 70% des données (2 383 individus) sont utilisés pour entraîner et évaluer le modèle cible. Ce sous-ensemble est lui-même divisé en 80% pour l'entraînement (1 907 individus) et 20% pour la validation (476 individus).
- Les 30% restants (1 021 individus) sont conservés comme données non vues par le modèle cible. Ces données servent à constituer les exemples « non-membres » pour l'évaluation finale de l'attaque.
- Le modèle d'attaque est entraîné uniquement à partir de jeux de données externes hétérogènes (Purchase, CIFAR, Adult, etc.), totalement indépendants des données de levure. Les modèles d'ombre associés produisent des vecteurs de sorties (postériors) qui alimentent un classifieur d'attaque généraliste.
- Pour tester ce classifieur sur le modèle cible, on forme ensuite un ensemble équilibré d'exemples :
  - « membres » : extraits de l'ensemble d'entraînement du modèle cible (parmi les 1 907 échantillons vus pendant l'apprentissage) ;
  - « non-membres » : extraits des 1 021 échantillons jamais vus par le modèle cible.

Aucune validation croisée systématique ni recherche exhaustive d'hyperparamètres n'ont été effectuées dans ce travail, car l'objectif principal n'était pas d'optimiser finement la performance prédictive de chaque modèle, mais d'évaluer leur comportement dans un scénario réaliste d'attaque. Nous avons toutefois procédé à un réglage manuel limité : pour le modèle cible, nous avons repris les hyperparamètres proposés par Chen *et al.* (2020) et vérifié, au moyen de quelques essais préliminaires (variation du nombre de filtres et du taux d'apprentissage), que les performances restaient stables. Pour les modèles d'ombre (régression logistique) et le modèle d'attaque (forêt aléatoire), nous avons conservé des configurations standards largement utilisées dans la littérature (régularisation L2 avec $C = 1{,}0$, profondeur maximale fixée à 2), après quelques tests exploratoires. Ce choix permet de contenir le

coût de calcul et se justifie par le fait que notre analyse porte avant tout sur la faisabilité et la significativité globale des attaques, plutôt que sur l'obtention de modèles parfaitement optimisés. De plus, une graine aléatoire fixe a été utilisée pour toutes les étapes comportant un tirage aléatoire (division des données, rééchantillonnage, entraînement), afin de garantir la reproductibilité des résultats.

3.8 Discussion des limites et biais potentiels

Bien que les deux méthodes soient complémentaires, elles ont aussi leurs limites, qu'il est important de mettre en évidence pour une évaluation approfondie de leur impact.

- Taille limitée de l'ensemble de données : le jeu de données génomiques utilisé repose sur quelques milliers d'individus seulement. Cette taille restreinte impose des compromis entre l'entraînement du modèle cible, la construction des jeux membres/non-membres et la création des modèles d'ombre. Elle augmente aussi le risque de surapprentissage et limite la validité externe des résultats, qui doivent être interprétés comme une preuve de faisabilité plutôt que comme une estimation définitive du risque pour des bases de données beaucoup plus volumineuses.
- Dépendance à une architecture de modèle spécifique : le modèle cible est un 1D-CNN particulier, dont l'architecture et les hyperparamètres ont été fixés en suivant Chen *et al.* (2020). Les résultats obtenus caractérisent donc avant tout la vulnérabilité de cette famille de modèles. D'autres architectures (réseaux plus profonds, régularisation plus forte, modèles linéaires ou ensembles) pourraient présenter un comportement différent vis-à-vis des attaques d'inférence d'appartenance, ce qui limite la généralisation immédiate des conclusions à l'ensemble des modèles utilisés en pratique sur des données biomédicales.
- Dépendance à la corrélation phénotypique : la première méthode repose sur l'hypothèse que les phénotypes auxiliaires utilisés pour entraîner les modèles d'ombre sont suffisamment corrélés avec le phénotype cible. Cependant, cette dépendance peut restreindre la capacité de généralisation si les phénotypes sélectionnés ne captent pas les mêmes signaux génétiques sous-jacents.
- Écart de distribution entre les jeux de transfert : dans la seconde méthode, les modèles d'ombre sont entraînés sur des ensembles de données externes, très différents du domaine génomique (images, textes, transactions, etc.). Ces écarts de distribution peuvent engendrer un décalage

entre les distributions de sorties des modèles d'ombre et celles du modèle cible, ce qui peut affecter les performances de l'attaque généralisée. Cependant, cette méthodologie est basée sur l'hypothèse que la différence entre les échantillons de membres et de non-membres se manifeste par le biais de motifs génériques dans les vecteurs de sortie (par exemple : la confiance du modèle, la dispersion des probabilités, etc.). Bien que les raisons derrière cela puissent différer en fonction des données, nos résultats expérimentaux montrent qu'ils sont suffisamment transférables pour permettre une attaque efficace, même lorsque les modèles d'ombre sont entraînés sur des domaines très éloignés du domaine cible. Cela suggère une certaine robustesse de l'attaque aux variations de distribution entre les domaines.

— Biais liés à l'utilisation de données de levure : les données utilisées pour entraîner le modèle cible proviennent de *Saccharomyces cerevisiae*, un organisme modèle unicellulaire. Cette approche permet un contrôle précis des variables génétiques et expérimentales. Cependant, il est encore nécessaire de démontrer la transférabilité des résultats à des données humaines. En effet, les génomes humains sont beaucoup plus complexes et variables, autant dans leurs interactions entre variants que dans l'effet de l'environnement.

— Limites computationnelles et structurelles : la méthodologie d'attaque généralisée impose un coût computationnel important lié à la manipulation de jeux de données hétérogènes et à l'entraînement de multiples modèles. De plus, elle suppose l'accès à des sorties probabilistes (softmax), ce qui pourrait ne pas être disponible dans certains services de prédiction.

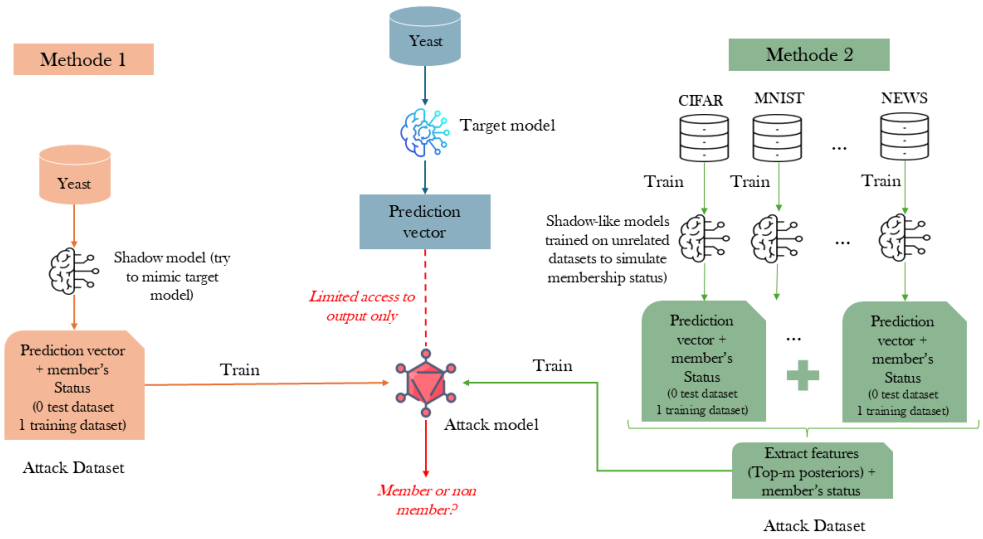

FIGURE 3.9 – Comparaison des deux approches d'attaque par inférence d'appartenance (MIA). La **Méthode 1** (à gauche) repose sur la construction d'un modèle d'ombre entraîné sur des phénotypes auxiliaires génétiquement corrélés à ceux du modèle cible (même distribution), afin de produire un vecteur de probabilité binaire servant à l'entraînement du modèle d'attaque. La **Méthode 2** (à droite) adopte une stratégie de transfert de connaissances généralisée, dans laquelle des modèles d'ombre sont entraînés sur des jeux de données hétérogènes (images, texte, données tabulaires) pour simuler des sorties membres/non-membres. Ces prédictions sont utilisées pour extraire des vecteurs de caractéristiques statistiques (top-m posteriors), qui servent ensuite à entraîner un modèle d'attaque. Le schéma intègre également une légende expliquant les rôles des composants (données, modèles, vecteurs de sortie, statut d'appartenance) pour faciliter la compréhension comparative.

3.9 Métrique de succès de l'attaque d'inférence d'appartenance

Étant donné que l'objectif principal des attaques par inférence d'appartenance est de déterminer si un échantillon a été utilisé lors de l'entraînement d'un modèle d'apprentissage automatique, il est essentiel d'évaluer correctement leur performance, non seulement pour mesurer leur efficacité, mais aussi pour estimer les risques de fuite de données sensibles dans des systèmes réels. Au fil des années, les recherches sur les MIA se sont concentrées sur l'exactitude (accuracy) comme indicateur d'évaluation. Cette métrique correspond à la proportion d'échantillons pour lesquels la prédiction du modèle d'attaque est correcte, qu'il s'agisse de membres (vrais positifs) ou de non-membres (vrais négatifs).

Pour mieux comprendre la manière dont cette métrique est calculée, il convient de présenter la matrice de confusion associée.

Réel / Prédit	**Non-membre**	**Membre**
Non-membre	TN	FP
Membre	FN	TP

La matrice de confusion est un outil fondamental en apprentissage automatique pour évaluer les performances d'un classificateur, qu'il soit binaire ou multiclasses. Elle permet de visualiser la répartition des prédictions du modèle par rapport aux classes réelles des échantillons.

Chaque cellule de cette matrice représente une combinaison possible entre la classe réelle d'un échantillon et la prédiction effectuée par le modèle d'attaque. Les valeurs attendues (classes réelles) sont affichées en ligne, tandis que les classes prédites figurent en colonne (Fergus et Chalmers, 2022).

Les vrais positifs (TP) correspondent aux échantillons effectivement membres, prédits comme tels.

Les faux positifs (FP) représentent les non-membres que le modèle a incorrectement classés comme membres.

Les faux négatifs (FN) sont des membres mal classés comme non-membres, tandis que les vrais négatifs (TN) désignent les non-membres correctement identifiés.

À partir de cette matrice, on peut définir l'exactitude (accuracy) comme la proportion d'échantillons

correctement classés par le modèle, indépendamment de leur appartenance :

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

Bien que l'exactitude fournisse une première indication sur la performance globale du modèle d'attaque, elle se révèle souvent insuffisante dans le cas particulier des attaques par inférence d'appartenance. En effet, cette métrique accorde la même importance aux prédictions correctes (TP et TN) et aux erreurs (FP et FN), sans prendre en compte leur impact différentiel. En effet, dans les bases de MIA, les faux positifs (FP), c'est-à-dire les échantillons non membres incorrectement identifiés comme membres, sont particulièrement problématiques. Cette erreur pourrait entraîner l'accusation injustifiée qu'une personne a été incluse dans un ensemble de données sensibles, ce qui constitue une violation grave de la vie privée.

Carlini *et al.* (2022) ont été les premiers à proposer une nouvelle approche d'évaluation, qui ne se fonde pas sur l'exactitude, mais sur deux métriques distinctes : le taux de vrais positifs (*True positive rate (TPR)*) et le taux de faux positifs (*False positive rate (FPR)*).

Le taux de vrais positifs (TPR) est défini comme :

$$\text{TPR} = \frac{TP}{TP + FN}$$

Et le taux de faux positifs (FPR) comme :

$$\text{FPR} = \frac{FP}{FP + TN}$$

Ces indicateurs permettent de mieux caractériser la capacité de l'attaque à distinguer les membres des non-membres, en mesurant séparément son efficacité (via le TPR) et son potentiel de nuisance (via le FPR). Intuitivement, une attaque efficace devrait maximiser le TPR, c'est-à-dire identifier correctement un grand nombre de membres, tout en minimisant le FPR, afin de limiter les fausses attributions. Le compromis TPR/FPR est entièrement représenté par la courbe ROC (*Receiver operating characteristic (ROC)*), qui trace le TPR en fonction du FPR pour toutes les valeurs possibles du seuil de décision. Cette courbe permet donc d'évaluer la performance globale de l'attaque, indépendamment d'un seuil arbitraire, et d'en dériver une mesure synthétique : l'aire sous la courbe (*AUC*). Plus l'AUC est proche de 1, plus l'attaque est discriminante. À l'inverse, une AUC proche de 0,5 indique un comportement aléatoire.

Par conséquent, l'évaluation des performances du modèle d'attaque dans cette étude repose principalement sur la courbe ROC, le couple TPR/FPR et l'aire sous la courbe (AUC), qui sont des indicateurs plus robustes et informatifs que la précision seule. En plus des métriques classiques, nous utilisons également deux mesures complémentaires : la précision (precision) et le rappel (recall).

La précision indique la proportion d'exemples prédits comme membres qui sont réellement membres (vrais positifs parmi tous les positifs prédits). Elle permet d'évaluer la fiabilité des prédictions positives du modèle d'attaque.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Le rappel, quant à lui, mesure la capacité du modèle à identifier correctement les membres réels (vrais positifs parmi tous les membres effectifs).

$$\text{Recall} = \frac{TP}{TP + FN}$$

Dans le contexte des attaques par inférence d'appartenance, les mesures telles que la précision et le rappel sont particulièrement utiles car elles permettent une évaluation plus nuancée des performances du modèle d'attaque. Elles mettent en évidence sa sensibilité aux faux positifs et aux faux négatifs. Cette approche est d'ailleurs employée dans les travaux de référence, comme celui de Shokri *et al.* (2017). Dans notre cas, nous avons construit les ensembles de données utilisés pour évaluer l'attaque de manière équilibrée, en incluant autant d'exemples membres et non membres. Cela garantit une évaluation équitable des mesures, telles que la précision et le rappel, sans biais introduit par un déséquilibre des classes.

3.10 L'environnement

Les expériences ont été menées dans un environnement Python 3.9, en s'appuyant sur plusieurs bibliothèques open source. Pour le prétraitement des données, nous avons utilisé NumPy (Harris *et al.*, 2020) et Pandas (McKinney *et al.*, 2010). Les modèles d'apprentissage supervisé, y compris les modèles d'ombre et les classificateurs d'attaque, ont été construits avec Scikit-learn (Pedregosa *et al.*,

2011). Le modèle cible, qui est un réseau de neurones convolutifs en dimension unique (1D-CNN), a été développé avec Keras Chollet (2015), en utilisant TensorFlow comme serveur (Abadi *et al.*, 2016a). Enfin, les métriques d'évaluation (accuracy, TPR, FPR, AUC), ainsi que les visualisations (matrices de confusion, courbes ROC), ont été générées à l'aide de Scikit-learn.metrics, Matplotlib (Hunter, 2007) et Seaborn (Waskom, 2021).

3.11 Conclusion

Dans ce chapitre, nous proposons un cadre méthodologique pour évaluer les risques pour la vie privée associés aux modèles d'apprentissage automatique utilisés avec des données génomiques. Ce cadre comprend les étapes clés suivantes : préparation des données, prétraitement, conception du modèle cible, création de modèles d'ombre et implémentation d'une attaque par inférence d'appartenance dans un scénario de type boîte noire.

Dans la première méthodologie, l'implémentation par modèle d'ombre classique, nous avons choisi des modèles ayant déjà fait leurs preuves dans des études antérieures en bio-informatique pour identifier un modèle d'ombre performant et généralisable. Ces modèles ont été choisis pour reproduire de manière réaliste le comportement du modèle cible. Parmi les modèles évalués, celui qui a permis la plus grande amélioration de la performance du modèle d'attaque a été choisi comme modèle d'ombre final. Ce modèle a servi à créer l'ensemble des données d'attaque en produisant des exemples étiquetés comme appartenant ou non à un groupe. Par la suite, une forêt aléatoire a été entraîné sur ces données pour mener l'attaque et identifier efficacement les échantillons appartenant au groupe cible.

En raison de la limitation du nombre d'échantillons disponibles pour entraîner un modèle d'ombre dans le domaine génomique, une deuxième méthodologie inspirée de l'article de Salem et al. a été adoptée pour implémenter l'attaque. Contrairement aux méthodologies classiques où le modèle d'ombre cherche à imiter le comportement du modèle cible, cette approche repose sur l'utilisation de plusieurs jeux de données publics et hétérogènes qui sont complètement différents du jeu de données utilisé pour le modèle cible. L'objectif n'est donc pas de copier exactement la logique du modèle cible, mais plutôt de capturer des schémas généraux permettant de distinguer les sorties associées aux membres et aux non-membres. Ainsi, un classifieur d'attaque généralisé, basé sur une forêt aléatoire,

a été entraîné à partir de ces données variées afin d'évaluer la robustesse du modèle cible face à des attaques d'inférence d'appartenance dans un scénario plus réaliste et moins dépendant des données.

Le chapitre suivant présentera les résultats expérimentaux obtenus en mettant en œuvre les méthodologies décrites plus tôt.

CHAPITRE 4
RÉSULTATS ET ANALYSE

Ce chapitre présente les résultats expérimentaux obtenus à partir des méthodologies décrites précédemment. Il vise à évaluer l'efficacité des attaques par inférence d'appartenance dans un contexte génomique réaliste. La performance de divers modèles (modèle cible, modèles d'ombre et modèle d'attaque) sera analysée selon plusieurs configurations.

Des expériences ont été menées sur des données réelles de levure, en utilisant des scénarios d'attaque de type boîte noire. Deux méthodes différentes ont été évaluées : la première consiste à créer un modèle d'ombre qui imite le comportement du modèle cible, tandis que la seconde met en œuvre une attaque par transfert de connaissances généralisée en utilisant des ensembles de données externes.

Les résultats sont analysés à l'aide de différentes métriques, telles que la précision, l'AUC, l'exactitude, le rappel et la matrice de confusion, afin de mesurer à la fois la puissance de détection de l'attaque et ses limites. Ce chapitre a également pour but de comparer les deux méthodologies et d'identifier les facteurs ayant le plus d'impact sur la réussite ou l'échec de l'inférence.

Dans l'article de référence, les résultats rapportés sont obtenus à partir d'un seul seed aléatoire. Afin de garantir une comparaison équitable, nous avons donc également présenté nos résultats principaux avec un seul seed. Cependant, pour évaluer la robustesse de notre méthode et vérifier son indépendance vis-à-vis du choix du seed, nous avons répété chaque expérience avec cinq seeds différents et rapporté les moyennes et écarts-types correspondants.

4.1 Évaluation du modèle cible

Le modèle cible, dont l'architecture a été décrite au chapitre précédent, a été entraîné sur un sous-ensemble des données génomiques de levure, en utilisant une classification binaire du phénotype « résistance au sulfate de cuivre ». Ce choix expérimental est directement inspiré de l'article de Chen *et al.* (2020), dans lequel ce phénotype est utilisé comme variable de sortie pour évaluer la vulnérabi-

lité des modèles d'apprentissage automatique face aux attaques par inférence d'appartenance.

L'objectif n'est pas d'optimiser les performances du modèle, mais simplement de vérifier qu'il atteint un niveau de précision adéquat pour jouer le rôle de cible dans l'attaque. À la fin de l'entraînement, le modèle affiche une précision supérieure à 95% sur les données d'entraînement, ainsi qu'une précision de validation qui se stabilise autour de 75 à 80%. Ce comportement est cohérent avec une situation de surajustement modéré, ce qui reste acceptable dans ce contexte expérimental, où l'accent est mis sur le comportement du modèle vis-à-vis de l'attaque plutôt que sur sa capacité de généralisation. Ces performances du modèle cible sont du même ordre de grandeur que celles rapportées par Chen *et al.* (2020), confirmant que notre configuration expérimentale est réaliste et comparable à l'état de l'art.

Pour le modèle cible, un total de 3404 échantillons a été utilisé. Ceux-ci ont été divisés en deux groupes :

— 2383 échantillons pour l'entraînement et le test du modèle cible ;

— 1021 échantillons comme jeu non vu (unseen) pour l'évaluation de l'attaque.

Le premier groupe a ensuite été subdivisé en :

— 1907 échantillons pour l'entraînement du modèle cible ;

— 476 échantillons pour son test.

4.2 Évaluation de l'attaque par modèle d'ombre

Dans cette section, nous examinons les résultats obtenus par l'attaque réalisée à l'aide de la première méthodologie, basée sur la construction d'un modèle d'ombre généralisé à partir de phénotypes statistiquement liés au phénotype cible. Contrairement à l'approche de référence de Chen *et al.* (2020), qui suppose que l'adversaire dispose de données étiquetées identiques à celles du modèle cible pour construire des modèles d'ombre, notre approche se veut plus réaliste : elle se fonde uniquement sur l'utilisation de données génétiques similaires, mais associées à des phénotypes différents. Pour choisir ces phénotypes auxiliaires, nous avons créé une matrice de corrélation entre 19 phénotypes mesurés sur un même groupe d'individus, en excluant le phénotype cible. Les phénotypes présentant les corrélations les plus fortes ont ensuite été utilisés pour entraîner un ou plusieurs modèles d'ombre, simulant ainsi le comportement du modèle cible sans jamais l'observer directement. Parmi les phénotypes auxi-

liaires évalués, seul Xylose a finalement été retenu pour construire le modèle d'ombre, car il présentait des performances supérieures lors de l'attaque, démontrant ainsi sa plus grande capacité à approximer le comportement du modèle cible.

TABLE 4.1 – Résumé des caractéristiques des modèles d'ombre et d'attaque dans la méthode 1

Caractéristiques	**Modèle d'ombre**	**Modèle d'attaque**
Données utilisées	Phénotype *Xylose* (distinct de la cible)	Sorties (scores) du modèle d'ombre pour l'entraînement, puis sorties du modèle cible pour l'évaluation
Nombre d'échantillons	4190	2042 (1021 membres, 1021 non-membres)
Proportion entraînement/test	50% entraînement, 50% test	– (tous les exemples utilisés pour l'entraînement ou l'évaluation)
Architecture du modèle	*Logistic Regression* (pénalité l_2, max_iter=1000)	*Random Forest* (profondeur max = 2)

Il est important de préciser que les 4190 profils utilisés pour entraîner le modèle d'ombre correspondent à l'ensemble complet des génotypes de levure disponibles après prétraitement (filtrage des valeurs manquantes et binarisation) pour le phénotype auxiliaire *Xylose*, distinct du phénotype cible (sulfate de cuivre). Ces profils ne servent pas à l'entraînement du modèle cible, et ne sont pas réutilisés pour définir les membres et non-membres de l'attaque finale. Pour l'attaque MIA proprement dite, les exemples « membres » sont exclusivement tirés du jeu d'entraînement du modèle cible, tandis que les « non-membres » proviennent du sous-ensemble *unseen* de 1021 individus qui n'ont jamais été vus pendant l'entraînement. Ainsi, les données employées pour apprendre le modèle d'ombre restent conceptuellement séparées du protocole d'évaluation de l'attaque, ce qui limite le risque de surapprentissage artificiel sur un ensemble de données particulier.

Les sorties probabilistes générées par le modèle d'ombre entraîné sur le phénotype Xylose ont été

utilisées pour constituer le jeu de données d'attaque. Chaque prédiction fournit un score de confiance pour chaque échantillon, permettant de l'associer à une étiquette binaire (membre ou non-membre) en fonction de son origine (jeu d'entraînement ou de test). Ces vecteurs de sortie, riches en information statistique, ont ensuite servi à entraîner un modèle d'attaque (forêt aléatoire) capable de distinguer les comportements typiques d'un échantillon vu par le modèle de ceux d'un échantillon inconnu.

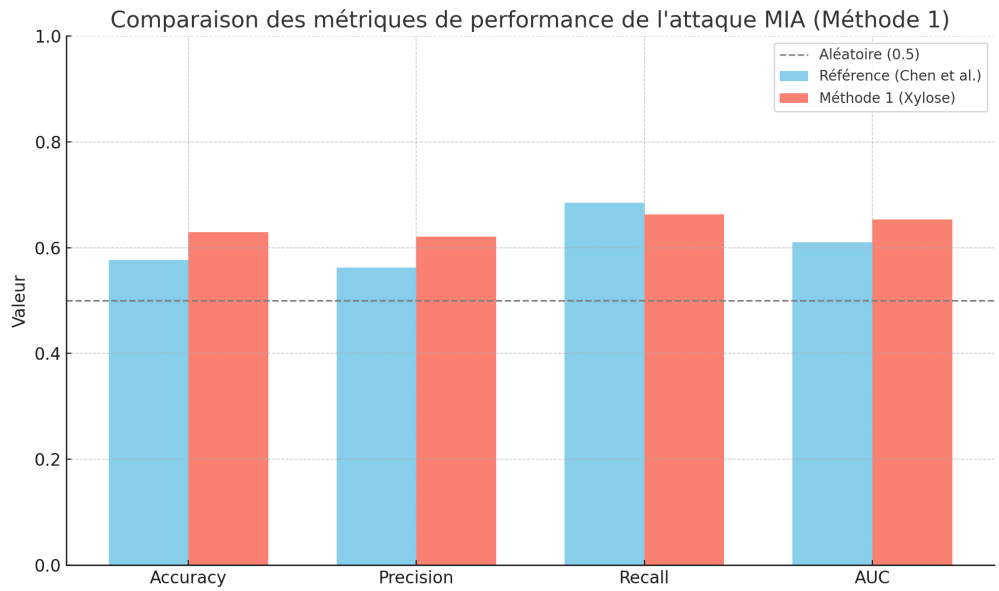

FIGURE 4.1 – Comparaison des performances de notre approche (modèle d'ombre entraîné sur le phénotype Xylose) avec celles de l'approche de Chen *et al.* (2020), selon quatre métriques classiques : exactitude, précision, rappel et AUC. On observe que notre méthode améliore systématiquement l'exactitude et l'AUC tout en réduisant le taux de faux positifs, au prix d'une légère baisse du rappel.

Comme le montre la Figure 4.1, pour le seed principal considéré, notre approche surpasse celle de Chen *et al.* (2020) sur la majorité des métriques : l'exactitude atteint 0,63 contre 0,58 pour la référence, la précision s'élève à 0,62 contre 0,56, et l'AUC passe de 0,615 à 0,655. Seul le rappel (TPR) reste légèrement inférieur à celui de la méthode de référence (0,663 contre 0,685), ce qui indique une très légère baisse dans la capacité à détecter tous les membres. Toutefois, cette différence est compensée par une réduction significative du taux de faux positifs.

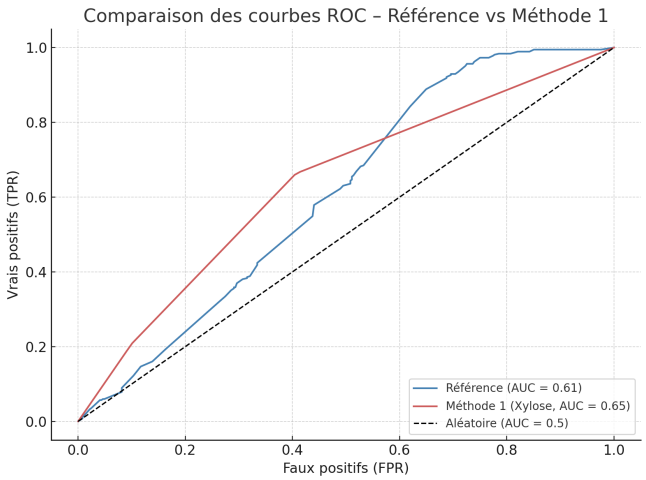

FIGURE 4.2 – Courbes ROC comparant la méthode de référence et notre approche (modèle d'ombre basé sur Xylose).

La Figure 4.2 présente les courbes ROC comparées. La courbe correspondant à notre méthode se situe systématiquement au-dessus de celle de Chen *et al.* (2020), ce qui traduit une meilleure capacité de séparation entre les échantillons membres et non-membres. L'amélioration de l'AUC confirme cette observation.

Les valeurs du Tableau 4.2 correspondent à la moyenne et à l'écart-type obtenus sur cinq exécutions indépendantes (cinq seeds aléatoires différents), ce qui permet de juger la stabilité de la méthode au-delà d'un seul seed.

TABLE 4.2 – Résumé des performances de l'attaque par modèle d'ombre corrélé (5 runs)

Métrique	**Moyenne**	**Écart-type**
Accuracy	0.6052	0.0171
Precision	0.6033	0.0161
Recall	0.6143	0.0282
F1-score	0.6086	0.0199
AUC	0.6127	0.0346

Dans l'ensemble, le tableau met en évidence une performance relativement stable et équilibrée pour la méthode basée sur les modèles d'ombre corrélés (Méthode 1). Les différentes métriques (exactitude, rappel, F1-score et AUC) se situent toutes autour de 60 %, avec de faibles écarts-types, ce qui confirme la robustesse de l'approche. Ces résultats suggèrent que l'utilisation de phénotypes biologiquement corrélés fournit une base efficace pour entraîner des modèles d'ombre pertinents dans le cadre de l'attaque MIA.

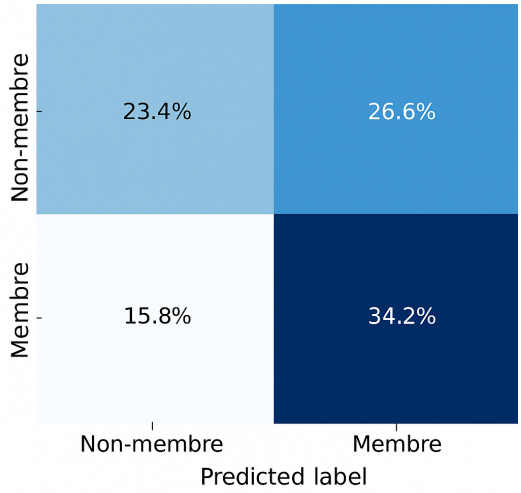

FIGURE 4.3 – Matrice de confusion – Méthode de référence

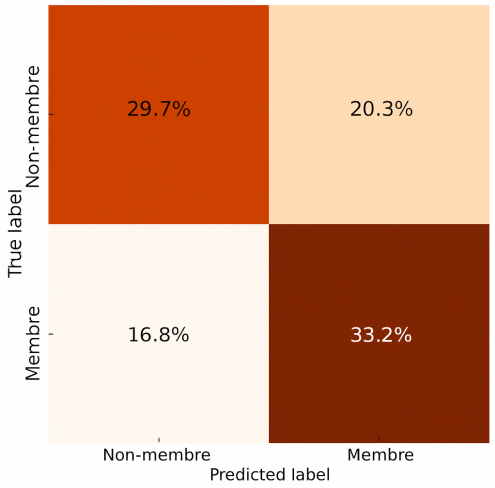

FIGURE 4.4 – Matrice de confusion – Notre méthode

L'analyse des matrices de confusion (Figures 4.3 et 4.4) révèle que notre approche atteint un compromis plus favorable entre la détection correcte des membres (TP) et la limitation des erreurs sur les non-membres (FP). Ce comportement est particulièrement avantageux dans les scénarios de type boîte noire où l'accès aux données est restreint. Bien que la détection correcte des membres (TP) soit légèrement réduite par rapport à la référence, la forte diminution du taux de faux positifs (FP) rend notre approche globalement plus robuste.

Méthode	**TPR**	**FPR**
Méthode de référence (Chen et al.)	**0,6848**	0,5326
Notre méthode (Xylose)	0,6631	**0,4058**

TABLE 4.3 – Comparaison des taux de vrais et faux positifs pour les deux méthodes

La Table 4.3 confirme que notre approche présente un taux de faux positifs nettement réduit (0,4058 contre 0,5326), ce qui diminue les alertes erronées et renforce la précision de l'attaque. Bien que légèrement moins performante en rappel, notre méthode reste plus fiable et plus sûre dans le contexte d'une attaque en boîte noire.

Ainsi, notre approche démontre qu'il est possible de mener des attaques d'inférence d'appartenance efficaces même dans des conditions réalistes et contraignantes, sans accès aux étiquettes ni à la structure du modèle cible. Cette contribution souligne l'urgence de développer des mécanismes de défense adaptés à ces nouvelles menaces en génomique computationnelle.

4.3 Évaluation de l'attaque par transfert de connaissances généralisée

Dans cette section, nous examinons de manière approfondie les résultats obtenus à l'aide de la méthode d'attaque par transfert de connaissances généralisée, inspirée des travaux de Salem *et al.* (2019). Contrairement à l'approche classique, où les modèles d'ombre sont construits dans des conditions similaires à celles du modèle cible, cette approche repose sur l'entraînement du modèle d'attaque à partir des sorties de modèles d'ombre hétérogènes, chacun étant formé sur un ensemble de données distinct et sans lien avec les données du modèle cible.

Les modèles d'ombre utilisés dans notre expérimentation couvrent plusieurs domaines et types de données : données tabulaires (Adult, Purchase), données textuelles (Newsgroups), données d'image (MNIST, CIFAR), ainsi que des données de localisation. Le rôle de chacun de ces modèles est uniquement de générer des échantillons membres et non-membres, à partir desquels sont extraits des vecteurs de probabilités. Ces vecteurs sont ensuite transformés en caractéristiques statistiques déri-

vées des probabilités de sortie (par exemple, les plus grandes probabilités parmi les classes). Dans le cas binaire du phénotype de levure, chaque exemple est en pratique résumé par la probabilité prédite pour la classe positive $p(y = 1 \mid x)$ (et sa probabilité complémentaire), ce qui constitue l'entrée du modèle d'attaque. Dans les jeux de données multi-classes utilisés pour entraîner certains modèles d'ombre externes (par exemple CIFAR-10/100 ou Purchase), nous retenons effectivement le top-k des probabilités de classes comme vecteurs de caractéristiques, avec $k = 10$ suivant Salem *et al.* (2019). En revanche, pour le modèle cible de levure qui est binaire, $k = 1$ et seul $p(y = 1 \mid x)$ (ainsi que sa probabilité complémentaire) est utilisé comme entrée du modèle d'attaque. Les vecteurs ainsi obtenus servent à alimenter un modèle d'attaque, en l'occurrence un classificateur de type forêt aléatoire, entraîné pour distinguer les membres des non-membres sans aucune connaissance préalable du modèle cible.

TABLE 4.4 – Résumé des caractéristiques des modèles d'ombre et d'attaque dans la méthode 2

Caractéristiques	**Modèles d'ombre**	**Modèle d'attaque**
Données utilisées	Résultats de modèles pré-entraînés sur des jeux hétérogènes (Purchase, Adult, CIFAR-10/100, MNIST, Location, News)	Sorties (postérieurs) du modèle cible sur les membres et non-membres
Nombre d'échantillons	500146	2042 (1021 membres, 1021 non-membres)
Proportion entraînement/test	50% entraînement, 50% test	Tous les exemples levure (2042) utilisés pour l'évaluation finale
Architecture du modèle	– (modèles pré-entraînés non spécifiés)	*Forêt aléatoire* (paramètres par défaut)
Format des entrées	Vecteurs de postérieurs (probabilités de classes) issus des modèles d'ombre	Probabilité de la classe positive prédite par le modèle cible (tâche binaire)

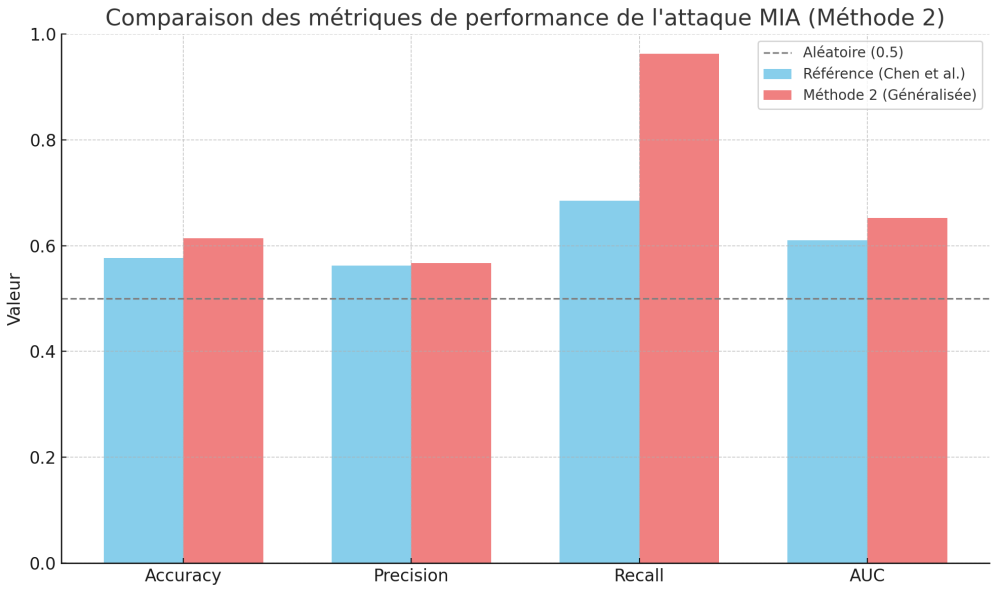

FIGURE 4.5 – Comparaison des performances globales entre la méthode de référence (boîte blanche) et notre approche généralisée (boîte noire) sur les métriques clés de l'attaque MIA. Malgré un accès plus limité au modèle cible, la méthode généralisée atteint une exactitude et une AUC supérieures, ainsi qu'un rappel nettement plus élevé, ce qui illustre la puissance du transfert de connaissances entre domaines.

L'objectif de cette évaluation est de mesurer la capacité de généralisation du modèle d'attaque à partir de sources hétérogènes. Cette propriété est essentielle dans des contextes réalistes, où l'adversaire ne dispose pas de données similaires à celles du modèle cible.

Les résultats illustrés dans la Figure 4.5 mettent en évidence une amélioration notable des performances obtenues avec notre méthode par rapport à la méthode de référence. Il faut noter que la méthode de référence se base sur un scénario en boîte blanche, dans lequel l'attaquant a un accès total à l'architecture et au poids du modèle cible. En revanche, notre approche s'inscrit dans un cadre réaliste d'attaque en boîte noire, où l'attaquant n'a accès qu'aux sorties du modèle cible. Cela rend notre méthode plus difficile, mais elle parvient tout de même à surpasser la référence sur plusieurs métriques. Les résultats obtenus sont comparés à ceux du modèle de référence proposé par Chen *et al.* (2020) à l'aide de plusieurs métriques : l'exactitude, la précision, le rappel, la courbe ROC et l'aire sous la courbe (AUC). La méthode généralisée atteint une exactitude de 61,4%, contre 57,6% pour

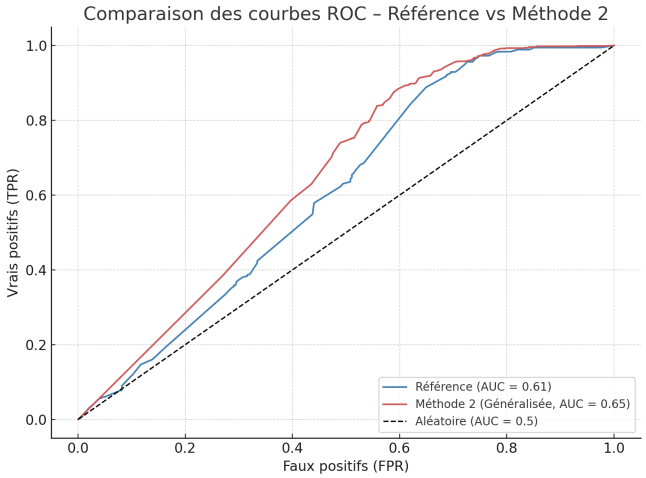

FIGURE 4.6 – Courbes ROC comparant la méthode de référence (boîte blanche) et notre approche (boîte noire). Notre méthode atteint une meilleure capacité de discrimination avec une AUC supérieure.

la méthode de Chen. Le rappel est particulièrement élevé (96,2% contre 68,4%), indiquant une capacité remarquable à identifier les membres. L'AUC s'améliore également, passant de 0,61 à 0,65, ce qui témoigne d'une meilleure séparabilité des classes. Bien que la précision reste relativement stable (autour de 56%), cela s'explique par une légère augmentation du taux de faux positifs, un compromis attendu dans une stratégie de rappel maximal.

Afin d'approfondir cette comparaison, chaque métrique est examinée individuellement ci-dessous à l'aide d'un graphique dédié et d'une interprétation contextuelle. L'exactitude mesure la proportion des prédictions correctes effectuées par le modèle d'attaque sur l'ensemble des échantillons. Bien qu'elle ne distingue pas les erreurs de type faux positif et faux négatif, elle offre une vue d'ensemble de la performance. Notre méthode atteint une exactitude de 61,4%, contre 57,6% pour la méthode de Chen, soit une amélioration absolue de 3,8 points. Cette progression est particulièrement significative compte tenu du contexte en boîte noire.

Pour tenir compte de la variabilité liée à l'aléatoire, chaque expérimentation a été répétée cinq fois avec des seeds différents (19122, 42, 1234, 2025 et 777). Les métriques présentées correspondent à la moyenne et à l'écart-type calculés sur ces cinq exécutions indépendantes. Le tableau 4.5 résume

les performances obtenues. Les valeurs du Tableau 4.5 correspondent à la moyenne et à l'écart-type obtenus sur cinq exécutions indépendantes (cinq seeds aléatoires différents), ce qui permet de juger la stabilité de la méthode au-delà d'un seul seed.

TABLE 4.5 – Résumé des performances de l'attaque généralisée (5 runs)

Métrique	**Moyenne**	**Écart-type**
Accuracy	0.59285	0.01479
Precision	0.55487	0.00830
Recall	0.93810	0.02434
F1-score	0.69727	0.01302
AUC	0.62753	0.01538

Les résultats du tableau 4.5 montrent que la méthode généralisée reste globalement stable et efficace sur cinq exécutions différentes. Le rappel élevé (93,8%) indique une forte capacité à détecter les membres, tandis que l'accuracy moyenne (59,3%) et l'AUC (62,7%) reflètent une bonne performance globale malgré un accès limité aux données du modèle cible. La faible variation (écarts-types modérés) confirme la robustesse de l'attaque face à l'aléa des initialisations.

Carlini *et al.* (2022) ont démontré que l'exactitude seule ne permet pas d'évaluer adéquatement l'efficacité réelle des attaques par inférence d'appartenance. Ils recommandent plutôt de combiner le taux de vrais positifs (TPR) et le taux de faux positifs (FPR), ainsi que de représenter les résultats sous forme de courbe *ROC*. Cette courbe montre le TPR en fonction du FPR pour différents seuils de décision, ce qui permet d'évaluer indépendamment de ce seuil la performance de l'attaque.

La courbe ROC obtenue montre que notre méthode d'attaque, bien qu'elle soit soumise à des contraintes plus strictes (boîte noire), surpasse la méthode de référence, qui a été développée dans un environnement plus ouvert (boîte blanche). En effet, la courbe correspondant à notre approche est toujours située au-dessus de celle de Chen *et al.* (2020). Cela démontre une meilleure capacité de distinction entre les membres et les non-membres.

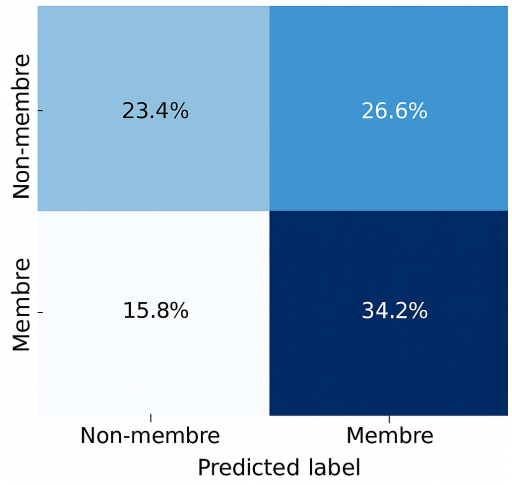

FIGURE 4.7 – Matrice de confusion – Méthode de référence

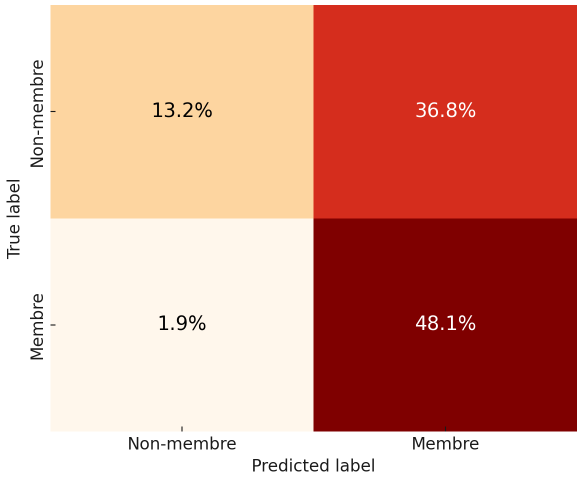

FIGURE 4.8 – Matrice de confusion – Notre méthode

Cette supériorité se manifeste par une courbe ROC systématiquement au-dessus de celle de la méthode de référence pour une large gamme de seuils de décision, ce qui montre qu'il existe des configurations où notre modèle peut atteindre un TPR plus élevé pour un niveau de FPR comparable. Cette caractéristique est cruciale dans les scénarios réels, où il est impératif de minimiser les faux positifs pour éviter des conclusions hâtives.

En outre, notre méthode atteint une aire sous la courbe (AUC) plus élevée, ce qui quantifie cette amélioration de la performance. Contrairement à la méthode de référence, qui nécessite un accès complet au modèle cible (structure et poids), notre approche ne nécessite que les sorties du modèle, tout en conservant une efficacité supérieure. Cette propriété revêt une importance particulière dans les scénarios réels, où la maîtrise du taux de faux positifs est essentielle afin d'éviter toute interprétation erronée.

Méthode	**TPR**	**FPR**
Méthode de référence (Chen et al.)	0,6848	**0,5326**
Notre méthode (généralisée)	**0,9628**	0,7356

TABLE 4.6 – Comparaison des taux de vrais et faux positifs pour les deux méthodes

Les taux de vrais positifs (TPR) et de faux positifs (FPR) permettent d'évaluer plus finement la capacité d'un modèle d'attaque à distinguer les membres des non-membres. Ces deux métriques sont fondamentales dans le cadre des attaques par inférence d'appartenance, notamment lorsqu'on souhaite minimiser les fausses alertes tout en maximisant la détection des échantillons réellement présents dans l'entraînement. Pour la méthode de référence (Chen et al.), le modèle atteint un TPR de 0,6848, indiquant qu'environ 68% des membres sont correctement identifiés. Le FPR est de 0,5326, ce qui signifie que plus de la moitié des non-membres sont incorrectement classés comme membres, entraînant un taux d'erreur non négligeable.

En revanche, notre méthode généralisée obtient un TPR remarquablement élevé de 0,9628, démontrant une capacité exceptionnelle à repérer les vrais membres. Toutefois, ce résultat s'accompagne d'un FPR plus élevé (0,7356), ce qui reflète une propension accrue à produire des faux positifs.

Ce compromis entre TPR et FPR est typique des approches cherchant à maximiser la sensibilité (recall) au détriment de la précision. Dans des scénarios de protection de la vie privée, une telle stratégie peut poser problème si elle aboutit à une surdétection des membres supposés.

Il convient aussi de souligner que la comparaison entre les deux méthodes se base sur des ensembles de données d'attaques de tailles différentes. La méthode de référence (Chen *et al.*, 2020) utilise un ensemble de données d'attaques créées dans un contexte contrôlé, généralement à partir d'un petit nombre d'échantillons provenant du même domaine que le modèle cible. En revanche, notre approche globale utilise un volume de données d'attaques générées à partir de nombreux modèles d'ombre entraînés sur des données hétérogènes.

Cette différence de taille ne constitue pas un biais, mais reflète une hypothèse réaliste dans laquelle l'adversaire peut accumuler davantage d'exemples d'attaque provenant de modèles d'ombre variés. Cela rend notre approche plus robuste et plus transférable dans des contextes réels, contrairement à l'attaque de Chen, très spécifique et dépendante du modèle cible.

TABLE 4.7 – Tableau récapitulatif des performances et caractéristiques des deux approches d'attaque par inférence d'appartenance proposées dans ce mémoire.

Critère	**Méthode 1 : Modèles d'ombre corrélés (Xylose)**	**Méthode 2 : Transfert de connaissances généralisé**
Type de données utilisées	Données réelles (levure), même distribution mais phénotype différent	Données externes (Adult, Purchase, MNIST, etc.), distributions hétérogènes
Accès au modèle cible	Boîte noire (sorties uniquement)	Boîte noire (sorties uniquement)
TPR	66,3%	**96,2%**
FPR	**40,6%**	73,6%
Exactitude	**63,0%**	61,4%
AUC	0,655	**0,657**
Robustesse au bruit / variabilité des données	Moyenne	Élevée
Dépendance au domaine	Moyenne (besoin d'un phénotype corrélé)	Faible (fonctionne avec données génériques)
Complexité de mise en œuvre	Moyenne (besoin d'analyse de corrélation)	Élevée (multidomaines, extraction statistique)

4.4 Synthèse comparative des deux méthodologies

Pour mieux comprendre les distinctions entre les deux approches d'attaque présentées dans ce chapitre, le tableau 4.7 propose une synthèse comparative de leurs caractéristiques et de leurs performances. En résumé, la méthode de transfert généralisé se distingue par une sensibilité accrue (TPR élevé), mais avec un taux de faux positifs plus élevé. En revanche, la méthode axée sur les phénotypes liés offre un meilleur équilibre entre précision et généralisation, tout en étant réaliste dans un contexte où l'accès direct aux données du modèle cible est restreint. Dans notre étude, nous avons utilisé l'exactitude (accuracy) comme principal indicateur d'évaluation des attaques. Cela nous a permis

de comparer directement nos résultats à ceux de l'étude de référence de Chen *et al.* (2020), qui adopte la même métrique dans un cadre génomique.

Cependant, à la lumière des recommandations de Carlini *et al.* (2022), nous reconnaissons que les métriques globales telles que l'accuracy ou l'AUC ne suffisent pas à elles seules à mesurer les risques de fuite. Ces auteurs soulignent qu'il est essentiel d'évaluer le taux de détection réel (TPR) à des taux de fausses alertes (FPR) faibles, afin de mieux cerner les menaces réelles pour la confidentialité.

Pour répondre à cette critique, nous avons intégré l'analyse conjointe des TPR et FPR, ainsi qu'un examen détaillé des matrices de confusion. Par exemple, dans la première méthodologie (modèles d'ombre corrélés), nous obtenons un TPR de 66,3% et un FPR de 40,6%, ce qui indique que l'attaque est capable d'identifier une proportion significative de membres, tout en maintenant un taux d'erreur modéré. En revanche, la seconde méthodologie (transfert de connaissances généralisé) atteint un TPR de 96,2%, mais au prix d'un FPR élevé (73,6%), traduisant un risque marqué de fausse classification des non-membres.

De plus, selon la définition proposée par Yeom *et al.* (2018), une attaque peut être jugée préoccupante dès lors que sa précision ou son rappel dépasse nettement le seuil de 50%. Dans notre cas, toutes les méthodes testées dépassent largement ce seuil, ce qui confirme leur faisabilité pratique, même dans un cadre contraint de type boîte noire et sur des données sensibles comme le génome.

Enfin, bien que nos scores d'exactitude ou d'AUC puissent paraître modérés (autour de 63–65%), leur interprétation doit être replacée dans le contexte. D'une part, ces niveaux sont considérés comme critiques par plusieurs auteurs (Yeom *et al.*, 2018), et d'autre part, même un AUC faible peut suffire à compromettre certains individus selon Carlini *et al.* (2022).

Bien que nous n'ayons pas explicitement testé la résistance des méthodes face à du bruit injecté dans les données, les différences observées dans les taux de faux positifs suggèrent que la méthode 2 pourrait être plus sensible à la structure interne du modèle cible. Comme l'ont montré Carlini *et al.* (2022) et Shokri *et al.* (2017), un FPR élevé peut signaler un mauvais alignement entre le modèle d'attaque et la frontière de décision réelle du modèle cible, traduisant souvent un surapprentissage sur

des signaux peu généralisables.

Dans certains cas limites, notamment pour des individus présentant un profil génétique marginal ou atypique par rapport à la distribution globale, la méthode 2 a tendance à les classer à tort comme membres. Cela indique que le modèle d'attaque peut confondre rareté et appartenance, soulevant ainsi des enjeux éthiques importants, notamment dans les contextes cliniques ou de recherche.

Ces observations confirment notre hypothèse de départ : une attaque MIA peut être rendue plus stable, plus réaliste et plus efficace si elle repose sur un proxy biologique pertinent, même lorsque l'accès direct aux données originales est restreint. En revanche, bien que la méthode généralisée offre un TPR supérieur, son taux de faux positifs remet en question l'idée qu'une généralisation complète soit toujours préférable dans des domaines sensibles comme la génomique.

Du point de vue opérationnel, ces résultats peuvent être interprétés en termes de scénarios concrets d'attaque. Dans un contexte expérimental en génomique, un attaquant pourrait par exemple chercher à vérifier si une souche particulière de levure, associée à un protocole de laboratoire spécifique ou à une collaboration industrielle, a été utilisée pour entraîner un modèle publié. Dans un contexte humain, un scénario analogue consisterait à déterminer si le génome d'un individu donné a contribué à un modèle clinique (par exemple pour prédire la réponse à un traitement). Dans les deux cas, une MIA réussie permet de relier un individu (ou une souche) à un jeu de données potentiellement sensible, ce qui constitue déjà une fuite d'information, même si le modèle ne révèle pas directement les génotypes complets.

Concrètement, un TPR de 60–65 % avec un FPR d'environ 40 % (méthode 1) signifie que, pour 100 individus réellement présents dans l'entraînement, l'attaquant peut en identifier correctement une soixantaine, au prix d'une quarantaine de faux positifs parmi les non-membres. À l'inverse, la méthode 2, avec un TPR proche de 96 % mais un FPR supérieur à 70 %, correspond à une stratégie de « surdétection » où presque tous les membres sont détectés, mais au prix d'un très grand nombre d'accusations erronées. Dans un cadre de recherche ou clinique, une telle configuration serait difficilement acceptable, car elle exposerait un grand nombre de participants non impliqués à un risque de ré-identification injustifiée.

Ces observations suggèrent que, dans la pratique, un attaquant rationnel adapterait son seuil de décision selon le contexte : soit en privilégiant un FPR plus faible (au détriment du rappel) lorsqu'il cherche quelques cibles avec une forte confiance, soit en acceptant un FPR plus élevé lorsqu'il dispose de mécanismes complémentaires de filtrage. Dans tous les cas, nos résultats montrent qu'un attaquant bien informé pourrait exploiter ces attaques dans des conditions réalistes, ce qui renforce l'importance d'intégrer des mécanismes de protection dans les pipelines d'analyse génomique.

En définitive, la méthode 1 s'aligne davantage avec notre objectif principal : démontrer la possibilité d'une attaque réussie sans accès direct aux mêmes données, à condition d'exploiter des structures biologiquement corrélées. Cette approche constitue un compromis pertinent entre faisabilité, réalisme et efficacité, tout en mettant en lumière les limites actuelles des défenses mises en place dans les systèmes d'analyse génomique.

Nos résultats s'inscrivent également dans le prolongement des attaques fondées sur le niveau de confiance ou la perte du modèle (Yeom *et al.*, 2018; Carlini *et al.*, 2022). Alors que ces approches se contentent souvent de se baser sur un seuil global appliqué à la probabilité prédite ou à la perte, nos deux méthodologies exploitent des informations supplémentaires : soit la structure biologique des phénotypes corrélés (méthode 1), soit la diversité de modèles d'ombre hétérogènes (méthode 2). Cela explique que nous atteignons des performances comparables, voire supérieures, à celles rapportées dans la littérature, malgré un cadre plus contraint de type boîte noire.

Par ailleurs, plusieurs travaux récents se sont intéressés à l'impact de techniques de régularisation ou de défense, telles que la régularisation adversariale de la perte d'appartenance (Nasr *et al.*, 2018) ou l'entraînement différentiellement privé (DP-SGD) (Abadi *et al.*, 2016b). Bien que ces mécanismes n'aient pas été explicitement évalués dans nos expériences, nos résultats fournissent une ligne de base pour de futures études qui combineraient nos scénarios d'attaque (phénotypes corrélés et transfert généralisé) avec ces défenses. Une question ouverte importante consiste à déterminer si ces méthodes restent efficaces lorsque l'attaquant n'a accès qu'aux sorties du modèle, comme dans nos scénarios de boîte noire.

CONCLUSION

Ce mémoire avait pour objectif principal de construire et d'évaluer un modèle d'attaque généralisable contre des modèles d'apprentissage automatique appliqués aux données génétiques, dans un cadre réaliste de boîte noire. Plutôt que de supposer un accès privilégié au modèle cible ou à ses données d'entraînement, l'étude explore la possibilité pour un adversaire d'inférer l'appartenance d'un échantillon en se basant uniquement sur des modèles d'ombre entraînés sur des données de distribution différente, voire des phénotypes biologiquement corrélés.

Notre travail apporte plusieurs contributions originales :

- la mise en œuvre concrète d'un cadre d'attaque MIA sur des données génétiques réelles (levure), avec simulation de modèle cible et construction de modèles d'ombre biologiquement informés ;
- l'adaptation du scénario MIA à des contraintes réalistes de confidentialité, sans accès aux données ni aux paramètres internes du modèle cible ;
- la proposition d'une approche de transfert généralisé pour les attaques MIA, permettant d'exploiter des modèles d'ombre hétérogènes et non alignés biologiquement.

Les résultats obtenus confirment la faisabilité et l'efficacité de ces attaques. La méthode 1 (basée sur un phénotype corrélé comme le xylose) obtient une exactitude de 63%, une précision de 62% et une AUC de 0,655, tout en maintenant un FPR raisonnable de 40,6%. La méthode 2 (transfert généralisé) atteint un TPR remarquable de 96,2%, mais au prix d'un FPR plus élevé (73,6%), mettant en lumière le compromis entre sensibilité et spécificité dans un cadre sans alignement biologique. Ainsi, si la méthode 2 illustre la puissance du transfert généralisé, la méthode 1 apparaît plus équilibrée en termes de compromis TPR/FPR. Dans une perspective pratique, la méthode 1 pourrait donc être privilégiée dans des contextes biomédicaux réels, où la minimisation des faux positifs est cruciale pour éviter des interprétations erronées ou des alertes inutiles.

Ce travail présente néanmoins plusieurs limites qui ouvrent la voie à des pistes de recherche futures. Tout d'abord, l'évaluation a été réalisée uniquement sur des données de levure, ce qui limite la portée des conclusions pour des contextes cliniques humains. Ensuite, la méthode de transfert généralisé souffre d'un taux de faux positifs élevé, ce qui la rend difficile à utiliser telle quelle dans des scénarios

biomédicaux sensibles. De plus, nous n'avons pas étudié l'impact de mécanismes de défense (par exemple DP-SGD, régularisation adversariale ou masquage des postérieurs), de sorte que la robustesse de nos attaques face à ces contre-mesures reste une question ouverte. Enfin, l'utilisation d'un seul type de modèle d'attaque (forêt aléatoire) ne permet pas de conclure sur l'optimalité architecturale de notre cadre.

Ces expériences montrent que la protection de la vie privée ne peut pas se limiter à restreindre l'accès aux modèles ou aux données : des informations résiduelles dans les sorties (scores de confiance) peuvent suffire à compromettre l'appartenance des individus. Il s'agit d'un signal d'alerte important pour les déploiements de systèmes d'IA dans le domaine biomédical. Elles démontrent également que des signaux d'appartenance peuvent être captés même dans des contextes de transfert entre domaines, confirmant la faisabilité d'attaques MIA dans un cadre strictement boîte noire.

Ce travail présente plusieurs perspectives concrètes :

— Tester les algorithmes avec des données humaines synthétiques, comme UK Biobank simulée, pour se rapprocher davantage des enjeux cliniques et éthiques réels ;
— Évaluer des mécanismes de défense tels que la confidentialité différentielle (DP-SGD), le masquage des postérieurs, ou la régularisation adversarielle ;
— Étendre l'analyse à d'autres types de données omiques (expression génique, épigénétique), afin d'évaluer la généralisabilité des attaques dans des espaces biologiques variés ;
— Explorer d'autres architectures pour le modèle d'attaque (réseaux neuronaux légers, modèles bayésiens calibrés) pour optimiser le compromis entre un TPR élevé et un FPR contrôlé ;
— Réduire la dépendance aux corrélations phénotypiques documentées en automatisant la sélection des proxys biologiques, ou en générant des jeux de données hybrides semi-synthétiques.

En conclusion, ce mémoire démontre que même dans un cadre strictement boîte noire, les attaques MIA peuvent réussir, notamment grâce à l'exploitation intelligente de signaux résiduels ou de proximités biologiques. L'approche par modèles d'ombre biologiquement corrélés constitue une contribution novatrice et efficace, applicable dans des contextes réels. Elle révèle que les corrélations naturelles présentes dans les données omiques, si elles ne sont pas encadrées, peuvent devenir des vulnérabilités

exploitables.

Sur le plan pratique, nos résultats soulignent plusieurs implications pour la conception et le déploiement de modèles en génomique. Dans un monde où les données génétiques sont de plus en plus partagées entre institutions, patients et systèmes d'intelligence artificielle, il est urgent d'intégrer des protections robustes dès la conception des modèles. D'autant plus que ces risques soulèvent également des enjeux légaux majeurs en matière de conformité aux réglementations internationales telles que le RGPD en Europe ou la HIPAA en Amérique du Nord. Ce travail s'inscrit dans cette logique : anticiper les attaques futures pour mieux défendre les individus, leur vie privée, et la confiance dans la recherche biomédicale.

BIBLIOGRAPHIE

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. *et al.* (2016a). {TensorFlow} : a system for {Large-Scale} machine learning. Dans *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 265–283.

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K. et Zhang, L. (2016b). Deep learning with differential privacy. Dans *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 308–318.

Abdollahi-Arpanahi, R., Gianola, D. et Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution*, *52*(1), 12.

Alnuaimi, A. F. et Albaldawi, T. H. (2024). An overview of machine learning classification techniques. Dans *BIO Web of Conferences*, volume 97, p. 00133. EDP Sciences.

Alzubi, J., Nayyar, A. et Kumar, A. (2018). Machine learning from theory to algorithms : an overview. Dans *Journal of physics : conference series*, volume 1142, p. 012012. IOP Publishing.

Ayday, E. (2016). Cryptographic solutions for genomic privacy. Dans *International Conference on Financial Cryptography and Data Security*, 328–341. Springer.

Ayday, E. et Humbert, M. (2017). Inference attacks against kin genomic privacy. *IEEE Security & Privacy*, *15*(5), 29–37.

Backes, M., Berrang, P., Humbert, M. et Manoharan, P. (2016). Membership privacy in microrna-based studies. Dans *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 319–330.

Bernau, D., Robl, J., Grassal, P. W., Schneider, S. et Kerschbaum, F. (2021). Comparing local and central differential privacy using membership inference attacks. Dans *IFIP Annual Conference on Data and Applications Security and Privacy*, 22–42. Springer.

Bloom, J. S., Kotenko, I., Sadhu, M. J., Treusch, S., Albert, F. W. et Kruglyak, L. (2015). Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. *Nature communications*, *6*(1), 8712.

Bonomi, L., Huang, Y. et Ohno-Machado, L. (2020). Privacy challenges and research opportunities for genomic data sharing. *Nature genetics*, *52*(7), 646–654.

Botta, V., Louppe, G., Geurts, P. et Wehenkel, L. (2014). Exploiting snp correlations within random forest for genome-wide association studies. *PloS one*, *9*(4), e93379.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Bu, D., Wang, X. et Tang, H. (2021). Haplotype-based membership inference from summary genomic data. *Bioinformatics*, *37*(Supplement_1), i161–i168.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A. et Tramer, F. (2022). Membership inference attacks from first principles. Dans *2022 IEEE symposium on security and privacy (SP)*, 1897–1914. IEEE.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U. *et al.* (2021). Extracting training data from large language models. Dans *30th USENIX security symposium (USENIX Security 21)*, 2633–2650.

Chang, H., Shejwalkar, V., Shokri, R. et Houmansadr, A. (2019). Cronus : Robust and heterogeneous collaborative learning with black-box knowledge transfer. *arXiv preprint arXiv :1912.11279*.

Chen, J., Wang, W. H. et Shi, X. (2020). Differential privacy protection against membership inference attack on machine learning for genomic data. Dans *BIOCOMPUTING 2021 : Proceedings of the Pacific Symposium*, 26–37. World Scientific.

Chen, X. et Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, *99*(6), 323–329.

Chollet, F. e. a. (2015). Keras. `https://github.com/fchollet/keras`.

Choquette-Choo, C. A., Tramer, F., Carlini, N. et Papernot, N. (2021). Label-only membership inference attacks. Dans *International conference on machine learning*, 1964–1974. PMLR.

Dabas, P., Kumar, D. et Sharma, N. (2017). Yeast genetics as a powerful tool to study human diseases. *Yeast Diversity in Human Welfare*, 191–214.

De Cristofaro, E. (2020). An overview of privacy in machine learning. *arXiv preprint arXiv :2005.08679*.

Erlich, Y. et Narayanan, A. (2014). Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, *15*(6), 409–421.

Fergus, P. et Chalmers, C. (2022). Applied deep learning. *Computational intelligence methods and applications*.

Fredrikson, M., Jha, S. et Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. Dans *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 1322–1333.

Gazestani, V. H. et Lewis, N. E. (2019). From genotype to phenotype : augmenting deep learning with networks and systems biology. *Current opinion in systems biology*, *15*, 68–73.

Grome, M. W. et Isaacs, F. J. (2021). Ztcg : viruses expand the genetic alphabet. *Science*, *372*(6541), 460–461.

Guo, T. et Li, X. (2023). Machine learning for predicting phenotype from genotype and environment. *Current Opinion in Biotechnology*, *79*, 102853.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. et Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, *339*(6117), 321–324.

Hagestedt, I., Humbert, M., Berrang, P., Lehmann, I., Eils, R., Backes, M. et Zhang, Y. (2020). Membership inference against dna methylation databases. Dans *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, 509–520. IEEE.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J. *et al.* (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362.

Holthouse, R., Owens, S. et Bhunia, S. (2025). The 23andme data breach : Analyzing credential stuffing attacks, security vulnerabilities, and mitigation strategies. *arXiv preprint arXiv :2502.04303*.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F. et Craig, D. W. (2008). Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, *4*(8), e1000167.

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S. et Zhang, X. (2022). Membership inference attacks on machine learning : A survey. *ACM Computing Surveys (CSUR)*, *54*(11s), 1–37.

Hunter, J. D. (2007). Matplotlib : A 2d graphics environment. *Computing in science & engineering*, *9*(03), 90–95.

Jayaraman, B. et Evans, D. (2019). Evaluating differentially private machine learning in practice. Dans *28th USENIX security symposium (USENIX security 19)*, 1895–1912.

Jia, J., Salem, A., Backes, M., Zhang, Y. et Gong, N. Z. (2019). Memguard : Defending against black-box membership inference attacks via adversarial examples. Dans *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 259–274.

Kansal, R. (2025). Rapid whole-genome sequencing in critically ill infants and children with suspected, undiagnosed genetic diseases : evolution to a first-tier clinical laboratory test in the era of precision medicine. *Children*, *12*(4), 429.

Katsara, M.-A., Branicki, W., Walsh, S., Kayser, M., Nothnagel, M., Consortium, V. *et al.* (2021). Evaluation of supervised machine-learning methods for predicting appearance traits from dna. *Forensic Science International : Genetics*, *53*, 102507.

LeCun, Y., Bottou, L., Bengio, Y. et Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Leino, K. et Fredrikson, M. (2020). Stolen memories : Leveraging model memorization for calibrated {White-Box} membership inference. Dans *29th USENIX security symposium (USENIX Security 20)*, 1605–1622.

Li, J., Li, N. et Ribeiro, B. (2021). Membership inference attacks and defenses in classification models. Dans *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, 5–16.

Liu, G., Wang, C., Peng, K., Huang, H., Li, Y. et Cheng, W. (2019). Socinf : Membership inference attacks on social media health data with machine learning. *IEEE Transactions on Computational Social Systems*, *6*(5), 907–921.

Liu, J., Huang, J., Zhou, Y., Li, X., Ji, S., Xiong, H. et Dou, D. (2022). From distributed machine learning to federated learning : A survey. *Knowledge and Information Systems*, *64*(4), 885–917.

Long, Y., Wang, L., Bu, D., Bindschaedler, V., Wang, X., Tang, H., Gunter, C. A. et Chen, K. (2020). A pragmatic approach to membership inferences on machine learning models. Dans *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, 521–534. IEEE.

Lourenço, V. M., Ogutu, J. O., Rodrigues, R. A., Posekany, A. et Piepho, H.-P. (2024). Genomic prediction using machine learning : a comparison of the performance of regularized regression, ensemble, instance-based and deep learning methods on synthetic and empirical data. *BMC genomics*, *25*(1), 152.

McGuire, A. L., Caulfield, T. et Cho, M. K. (2008). Research ethics and the challenge of whole-genome sequencing. *Nature Reviews Genetics*, *9*(2), 152–156.

McKinney, W. *et al.* (2010). Data structures for statistical computing in python. *SciPy*, *445*(1), 51–56.

McMahan, B., Moore, E., Ramage, D., Hampson, S. et y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Dans *Artificial intelligence and statistics*, 1273–1282. PMLR.

Muhamedyev, R. (2015). Machine learning methods : An overview. *Computer modelling & new technologies*, *19*(6), 14–29.

Nasr, M., Shokri, R. et Houmansadr, A. (2018). Machine learning with membership privacy using adversarial regularization. Dans *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 634–646.

Oestreich, M., Chen, D., Schultze, J. L., Fritz, M. et Becker, M. (2021). Privacy considerations for sharing genomics data. *EXCLI journal*, *20*, 1243.

Orgogozo, V., Morizot, B. et Martin, A. (2015). The differential view of genotype–phenotype relationships. *Frontiers in genetics*, *6*, 179.

Papernot, N., McDaniel, P., Sinha, A. et Wellman, M. P. (2018a). Sok : Security and privacy in machine learning. Dans *2018 IEEE European symposium on security and privacy (EuroS&P)*, 399–414. IEEE.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K. et Erlingsson, Ú. (2018b). Scalable private learning with PATE. Dans *International Conference on Learning Representations (ICLR)*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. *et al.* (2011). Scikit-learn : Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Poswal, A. M. et Saini, A. K. (2017). Yeast as a model system to study human diseases. *Metabolic Engineering for Bioactive Compounds : Strategies and Processes*, 209–220.

Rahman, M. M., Arshi, A. S., Hasan, M. M., Mishu, S. F., Shahriar, H. et Wu, F. (2023). Security risk and attacks in ai : A survey of security and privacy. Dans *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1834–1839. IEEE.

Rigaki, M. et Garcia, S. (2023). A survey of privacy attacks in machine learning. *ACM Computing Surveys*, *56*(4), 1–34.

Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M. et Backes, M. (2019). Ml-leaks : Model and data independent membership inference attacks and defenses on machine learning models. Dans *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS)*. The Internet Society.

Schübeler, D. (2015). Function and information content of dna methylation. *Nature*, *517*(7534), 321–326.

Sehrawat, S., Najafian, K. et Jin, L. (2023). Predicting phenotypes from novel genomic markers using deep learning. *Bioinformatics Advances*, *3*(1), vbad028.

Shejwalkar, V. et Houmansadr, A. (2021). Membership privacy for machine learning models through knowledge transfer. Dans *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 9549–9557.

Shokri, R., Stronati, M., Song, C. et Shmatikov, V. (2017). Membership inference attacks against machine learning models. Dans *2017 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.

Skelly, D. A., Merrihew, G. E., Riffle, M., Connelly, C. F., Kerr, E. O., Johansson, M., Jaschob, D., Graczyk, B., Shulman, N. J., Wakefield, J. *et al.* (2013). Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome research*, *23*(9), 1496–1504.

Song, C. et Raghunathan, A. (2020). Information leakage in embedding models. Dans *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, 377–390.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica*, *24*(1), 12–18.

Tian, Z., Cui, L., Liang, J. et Yu, S. (2022). A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, *55*(8), 1–35.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K. et Ristenpart, T. (2016). Stealing machine learning models via prediction {APIs}. Dans *25th USENIX security symposium (USENIX Security 16)*, 601–618.

Truex, S., Liu, L., Gursoy, M. E., Yu, L. et Wei, W. (2019). Demystifying membership inference attacks in machine learning as a service. *IEEE transactions on services computing*, *14*(6), 2073–2089.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. et Yang, J. (2017). 10 years of gwas discovery : biology, function, and translation. *The American Journal of Human Genetics*, *101*(1), 5–22.

Wang, R., Li, Y. F., Wang, X., Tang, H. et Zhou, X. (2009). Learning your identity and disease from research papers : information leaks in genome wide association study. Dans *Proceedings of the 16th ACM conference on Computer and communications security*, 534–544.

Wang, S., Jiang, X., Singh, S., Marmor, R., Bonomi, L., Fox, D., Dow, M. et Ohno-Machado, L. (2017). Genome privacy : challenges, technical approaches to mitigate risk, and ethical considerations in the united states. *Annals of the New York Academy of Sciences*, *1387*(1), 73–83.

Waskom, M. L. (2021). Seaborn : statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Wetterstrand, K. A. (2023). Dna sequencing costs : Data from the nhgri genome sequencing program (gsp). `https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data`. National Human Genome Research Institute (NHGRI). Consulté le 10 nov. 2025.

Wong, K.-C., Zhang, J., Yan, S., Li, X., Lin, Q., Kwong, S. et Liang, C. (2019). Dna sequencing technologies : sequencing data protocols and bioinformatics tools. *ACM Computing Surveys (CSUR)*, *52*(5), 1–30.

Wright, F. et Fessele, K. (2017). Primer in genetics and genomics, article 5—further defining the concepts of genotype and phenotype and exploring genotype–phenotype associations. *Biological research for nursing*, *19*(5), 576–585.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. et Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, *25*(6), 714–721.

Xue, M., Yuan, C., Wu, H., Zhang, Y. et Liu, W. (2020). Machine learning security : Threats, countermeasures, and evaluations. *IEEE Access*, *8*, 74720–74742.

Yang, D., Zhang, D., Zheng, V. W. et Yu, Z. (2014). Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics : Systems*, *45*(1), 129–142.

Yeom, S., Giacomelli, I., Fredrikson, M. et Jha, S. (2018). Privacy risk in machine learning : Analyzing the connection to overfitting. Dans *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, 268–282. IEEE.

Zerhouni, E. A. et Nabel, E. G. (2008). Protecting aggregate genomic data. *Science*, *322*(5898), 44–44.

Zhang, X., Fang, C. et Shi, J. (2021). Thief, beware of what get you there : Towards understanding model extraction attack. *arXiv preprint arXiv :2104.05921*.

Zhao, J., Bodner, G. et Rewald, B. (2016). Phenotyping : using machine learning for improved pairwise genotype classification based on root traits. *Frontiers in plant science*, *7*, 1864.

Zhou, Z.-X. et Kunkel, T. A. (2022). Extrinsic proofreading. *DNA repair*, *117*, 103369.