UNIVERSITÉ DU QUÉBEC À MONTRÉAL

AMÉLIORATION DE L'ÉQUITÉ PAR DES APPROCHES DE PRÉTRAITEMENT ADVERSÉRIELLES

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN INFORMATIQUE

PAR

ROSIN CLAUDE NGUEVEU

JUIN 2023

REMERCIEMENTS

De nombreuses personnes m'ont entouré durant cette grande aventure de ma vie qu'est la réalisation de ma thèse. Je ne saurai toutes les remercier dans cet espace assez limité, car le faire demanderait l'écriture d'un livre entier.

Premièrement, il me tient à coeur de remercier mon directeur de thèse, Sébastien Gambs, qui m'a accompagné tout au long de cette aventure. Son appui, ses encouragements, ses conseils, son soutien permanent dans ma recherche et la direction de celle-ci ont permis que je puisse garder l'esprit ouvert, réaliser les travaux de recherches passionnants sans être totalement dispersé dans ce domaine. Je tiens aussi à remercier mon codirecteur Alain Tapp, qui a dirigé ma recherche et m'a accompagné durant sa réalisation.

Je ne saurai suffisamment remercier ma famille, mon père, ma mère, Aude, Lorys, Sandra, Yann, Père Alain, qui ont toujours été avec moi dans les moments difficiles, comme dans les meilleurs moments. Leurs encouragements, aides de multiples façons, ont permis à ce que cette thèse puisse être réalisée et complétée, tout en me permettant de garder la tête hors de l'eau.

Ma thèse a aussi pu être amenée à son achèvement grâce à l'aide de mes proches, Ulrich (a.k.a shisho), Daniella L., David, Daniella N., Steve, Polux, Maguy, Sr. Estelle, Landry, Gisèle, Yeshna, Bahvvya, Pritamsing (Coach), Rachelle pour toutes leurs paroles de sagesses, leurs prières et leur soutien moral.

Un merci spécial est adressé à Antoine, Louis, Sylvain (a.k.a Papounet) et Marc-Olivier, pour les échanges constructifs, les conseils, les différentes assistances techniques et matérielles, les conseils de rédactions, de présentations et de direction de recherche.

Avant de conclure, un grand merci à tous les membres anciens et nouveaux de l'équipe de recherche PrivSec, pour la bonne ambiance dans le laboratoire, les échanges instructifs et même drôles. En particulier, je remercie tous les stagiaires sous ma supervision, pour l'apprentissage de l'humilité et de l'encadrement d'autres potentiels chercheurs.

Enfin, je remercie grandement toutes les personnes qui, de près ou de loin, ont collaboré à cette thèse et m'ont soutenue de diverses façons pour que je puisse terminer ma recherche de doctorat.

Merci à toute personne qui lira ce document de recherche, fruit de plusieurs années de travail.

DEDICACE

AVANT-PROPOS

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

Symbole	Signification
ML	Apprentissage Automatique (*Machine learning*)
R	Ensemble de données
N	Nombre d'individus (lignes) présent dans l'ensemble de données
Bs	Taille d'un lot d'entraînement
d	Nombre d'attributs (dimension) du jeu de données
k	Nombre maximal d'attributs sensibles
r_i $(i \in \{1, \ldots, N\})$	Caractéristique de l'individu i de l'ensemble R
S_j	j^{me} attribut sensible
Y	Attribut de décision
A	Attributs qui ne sont ni sensibles, ni de décision
Z	Représentation de l'information dans un espace latent ou espace non-interprétable.
X	Bruit aléatoire utilisé en entrée des modèles
BER	Taux d'erreur équilibré (*Balanced Error Rate*)
$SAcc$	Précision de la prédiction de l'attribut sensible (*Sensitive attribute Accuracy*)
Acc_T	Exactitude de prédiction de l'attribut T
$DemoParity$	Parité démographique (*Demographic Parity*)
$EqGap$ ou $EqGap_y$	Égalité des chances
EOD ou $EqGap_{y=1}$	Égalité des opportunités
yNN	Consistence (*Consistency*)
Fid	Fidelité
$Diversité$	Diversité
AttGAN	Facial Attribute Editing by Only Changing What You Want (He, Zuo, Kan, Shan et al., 2019)
WGAN	Wasserstein generative adversarial networks (Arjovsky, Chintala et Bottou, 2017)
GANSan	Generative Adversarial Network Sanitization (Aïvodji, Bidet, Gambs, Ngueveu et al., 2021)
DIRM	Disparate Impact Remover (Feldman, Friedler, Moeller, Scheidegger et al., 2015)
FM	FairMapping
FM2D	FairMapping version avec 2 discriminateurs
$P.M$	Transformation des données du groupe protégé uniquement.
$A.M$	Transformation de tous les individus de l'ensemble de donnée.
$Y.O$	Processus de transformation réalisé sans prendre en compte l'attribut de décision.
$Y.M$	Transformation des données en prenant en compte la décision.
C	Classifieur chargé de prédire l'attribut sensible original sur des données non modifiées
D	Discriminateur chargé de prédire l'attribut sensible à partir de données modifiées
D_{std}	Discriminateur chargé de différentier les données réelles des données générées
G_{Enc}	Generateur Encodeur
G_{Dec}	Generateur Décodeur
G_{std}	Generateur standard, transformant N en Z
S_{an}	Assainisseur
λ_C	Coefficient de Classification
λ_D	Coefficient de Protection
$\lambda_{D_{std}}$	Coefficient GAN
λ_R	Coefficient de l'opération d'identité
α	Coefficient gérant le compromis entre la protection de l'attribut sensible et la reconstruction des données dans $GANSan$
D_{tv} ou D_∞	Mesures de distances entre individus
MILP	Programmation Linéaire en entier multiple
ACC	Exactitude de prédiction
AUC	Aire sous la courbe
FPR	Taux de Faux positifs
FNR	Taux de Faux négatifs
DD	Discrimination Directe
ID	Discrimination Indirecte
TE	Effet total
CE	Effet contrefactuel
DI	Effet disparate
DP	Confidentialité différentielle
BS	Score de Brier
RAG	Réseaux Adversariaux Génératifs
WGAN	Wasserstein Réseaux Adversariaux Génératifs
LFR	Apprentissage de représentation équitable (Zemel, Wu, Swersky, Pitassi et al., 2013)
ALFR	Apprentissage de représentation équitable par des adversaires (Edwards et Storkey, 2015)
MUBAL	Atténuation de biais par de l'apprentissage adversariel (Zhang, Lemoine et Mitchell, 2018)
LATR	Apprentissage de représentation équitable et transférable (Madras, Creager, Pitassi et Zemel, 2018)

RÉSUMÉ

L'apprentissage automatique est un domaine de recherche en pleine expansion, dont les applications sont nombreuses et impactent de nombreux aspects de notre vie. Par exemple, des modèles d'apprentissage sont déployés dans nos téléphones intelligents ou encore dans des systèmes d'aide à la décision dans le domaine financier ou juridique. Bien que ces outils aient un succès important et peuvent améliorer la qualité de vie, ils ne sont pas exempts de défauts. Ainsi, ces systèmes peuvent agir comme des miroirs et reproduire des biais sociétaux qu'ils ont appris, voir même les amplifier de manière systématique. Afin de pallier ces problèmes, de nombreuses recherches ont été menées ces dernières années pour obtenir des systèmes capables d'apprendre automatiquement tout en faisant abstraction des biais non désirables qui pourraient exister dans les données. Dans cette thèse, nous décrivons dans un premier temps les sources potentielles de biais. Nous explorons ensuite les méthodes et les mécanismes de quantification de la discrimination existant dans la littérature, et nous présentons quelques approches permettant de remédier aux biais dans l'apprentissage automatique. En plus de ces contributions, nous introduisons une catégorisation des méthodes d'amélioration de l'équité afin d'identifier les différents challenges et opportunités dans ce domaine.

Dans notre deuxième recherche, nous proposons l'approche d'amélioration de l'équité, GANSan, basée sur des systèmes de modélisation de distributions tels que les réseaux adversériaux génératifs. Nous conduisons une analyse étendue de l'impact de notre approche au travers de divers scénarii potentiels. Nous analysons ensuite les comportements de GANSan afin de comprendre différents phénomènes que nous avons pu observer, tels que l'amélioration des performances de prédiction d'une tâche. Ces analyses nous permettent d'établir les limites de notre approche dans la protection d'informations sensibles, ainsi que de montrer et d'expliquer le phénomène d'alignement des distributions.

Inspirée de notre deuxième chapitre, notre troisième chapitre propose l'approche DYSAN, qui est une amélioration de GANSan appliquée dans le contexte médical, particulièrement dans le domaine du *soi quantifié*. Cette nouvelle approche est personnalisable aux données des personnes, affiche des performances supérieures à l'état de l'art sur plusieurs jeux de données contenant des signaux de capteurs, et impacte légèrement l'autonomie de l'appareil sur lequel il est déployé. Ainsi, les utilisateurs peuvent partager leurs données tout en préservant leurs informations sensibles.

Enfin, notre dernière contribution majeure consiste en notre approche FairMapping, dont le but est de transférer les propriétés d'un premier groupe à un second, pour faire profiter au second les privilèges du premier, et pour mitiger les effets de la discrimination. Notre approche met en oeuvre le transport optimal pour transformer les données. FairMapping maintient l'aspect réaliste des données, réduit les risques de discrimination par la protection de l'attribut sensible, facilite l'interprétation et l'utilisation des données transformées dans des situations critiques et limite les modifications à faire sur un environnement existant pour utiliser les données modifiées. Les expériences réalisées sur deux ensembles de données permettent de valider nos hypothèses.

INTRODUCTION

De nos jours, l'adoption de l'apprentissage automatique se fait de manière de plus en plus importante, à tel point qu'il est présent dans presque toutes les sphères de la société. Ainsi, il se retrouve aussi bien dans la sphère privée avec l'émergence des téléphones intelligents, des assistants personnels (dont le nombre d'équipements qui en sont dotés dépasse le seuil de sept milliards (Sujay V., 2021)), des objets connectés (Alemdar, Tunca et Ersoy, 2015), que dans la sphère et l'administration publique (Agarwal, 2018 ; Anastasopoulos et Whitford, 2019 ; Sidey-Gibbons et Sidey-Gibbons, 2019 ; Stilgoe, 2018 ; Veale et Brass, 2019).

Par exemple, selon le centre de recherche Pew, près de la moitié des Américains (États-Unis) ont recours aux assistants personnels tels que SIRI ou Google Home, etc.) (CENTER, 2017). Dans la sphère publique, certains acteurs étatiques ou certaines juridictions ont recours à de tels systèmes pour assister les juges dans la prise de décision (Chen, 2021) ou encore faciliter les travaux administratifs.

« L'apprentissage automatique peut être définie comme le sous-domaine de l'*intelligence artificielle* qui s'intéresse à conférer aux ordinateurs une capacité d'apprentissage sans programmation explicite. Il s'agit ici de répondre à la question de savoir comment construire des programmes qui seraient en mesure de s'améliorer automatiquement à partir des connaissances obtenues au travers de différentes expérimentations. » (Samuel, 1967)

Arthur Samuel a proposé cette définition dans ses travaux portant sur la conception d'un programme capable de jouer aux dames, à partir desquels il conçut un mécanisme d'apprentissage permettant d'obtenir des performances supérieures à celles d'une personne choisie de manière aléatoire (Samuel, 1967). De manière plus formelle, un programme informatique apprend d'une expérience **E** relativement à un ensemble de tâches **T** et une mesure de performance **P** si les performances obtenues en réalisant les travaux **T** évalués avec **P** s'améliorent avec l'expérience **E** (Mitchell, 1997).

Typiquement, dans le cas du programme d'Arthur Samuel conçu pour le jeu de dames, le programme acquiert de l'expérience **E** en jouant contre lui-même, la tâche **T** consiste au jeu de dames et la mesure

de performance est la probabilité de remporter la victoire contre un futur adversaire (Samuel, 1967). Dans un autre contexte plus critique tel que la justice prédictive, la tâche **T** consiste en l'inférence du risque de récidive des détenus en se basant sur leurs caractéristiques. Le programme réalise des inférences sur le risque de récidive des détenus, qui est ensuite comparé à des valeurs connues d'avance. La différence entre le risque prédit et celui connu d'avance constitue l'expérience **E** à partir de laquelle le programme s'améliore. L'objectif est d'obtenir un risque prédit très proche ou égal à celui connue d'avance.

Cette modification de l'architecture du programme en fonction de l'expérience acquise constitue la phase d'*entraînement* du programme. En fonction des spécifications données et du résultat souhaité par la personne utilisatrice du programme, on peut distinguer plusieurs formes d'apprentissage, parmi lesquels l'apprentissage non-supervisé et l'apprentissage supervisé (Ayodele, 2010). En apprentissage supervisé, les éléments en entrée du programme et les résultats attendus sont spécifiés. Un algorithme d'apprentissage est ensuite exécuté pour construire un programme appelé modèle. Ce modèle appris devrait être en mesure de produire ces résultats attendus à partir des éléments en entrées spéficiés, et par la suite, il devrait être en mesure de produire des résultats appropriés à partir de nouveaux éléments d'entrée n'ayant pas servi à l'acquisition d'expérience. En d'autres termes, le modèle devrait-être capable de généraliser sur les nouveaux éléments d'entrées. Dans le cas de l'apprentissage non-supervisé, les résultats attendus ne sont pas spécifiés. Le modèle doit apprendre de lui même à extraire, à partir des données d'entrées, les informations nécessaire à la réalisation de la tâche.

En raison de cette construction de programme sans programmation explicite, l'apprentissage automatique permet de s'abstraire de contraintes nécessitant la définition formelle et explicite de la méthode de résolution d'un problème donné (c'est-à-dire de sa résolution exacte), pour se focaliser sur l'identification de la tâche à résoudre. En effet, concevoir un algorithme permettant de résoudre de manière exacte des problèmes réels en toute circonstance peut s'avérer difficile, voir impossible. Par exemple, concevoir des algorithmes de détection de visage au sein d'une image serait difficilement réalisable en identifiant explicitement les « pixels » correspondant aux différents visages en question. En effet, la position des visages (et des personnes) au sein de l'image, la forme, la couleur et autres caractéristiques peuvent varier significativement d'une image à une autre empêchant ainsi une codification explicite.

Les algorithmes d'apprentissage apportent donc une flexibilité dans la réalisation de la tâche, per-

mettent d'entraîner des modèles qui automatisent le choix des caractéristiques descriptives les plus utiles, peuvent obtenir de meilleures performances en termes de précision de la prédiction, et surtout offrent une facilité de réutilisation de la solution développée (O'Mahony, Campbell, Carvalho, Harapanahalli et al., 2019). La prise en compte d'éléments ou de caractéristiques ambiguës, difficilement exprimables mathématiquement (ou bien oralement) est ainsi facilitée par ces algorithmes, qui complètent en conséquence les méthodes numériques pour la résolution de problèmes.

De même, l'apprentissage automatique permettrait à la machine d'incorporer (voire de simuler) le raisonnement humain pour la réalisation de tâches précises (Carbonell, Michalski et Mitchell, 1983). Ce raisonnement est en partie capturé par un ensemble de données et les corrélations existantes, qui reflètent l'état du monde réel (au moment où elles ont été collectées) sans toutefois expliciter la cause de ces corrélations ou le raisonnement ayant permis d'associer entre elles chacune des caractéristiques descriptives des données.

Malgré son énorme succès, l'utilisation de l'apprentissage automatique ne se fait pas sans risques. En effet, ces systèmes intelligents apprennent à partir d'observations - qui sont matérialisées par les données, et peuvent reproduire et décupler de manière systématique des comportements indésirables existant dans le monde réel comme des différences de traitement entre personnes (Arsenault, 2022). Ils peuvent aussi apprendre des corrélations non souhaitées dont un humain n'aurait pas tenu compte pour exécuter une tâche (Yudkowsky et al., 2008) ou encore ne pas suffisamment apprendre de corrélations par manque de données (Buolamwini et Gebru, 2018). Par exemple, on peut mentionner l'outil de recherche et d'annotation d'images de Google, qui fut critiqué en 2015 parce qu'il avait assigné des étiquettes de gorilles à des personnes de couleurs (Vincent, 2018). De même, dans le domaine judiciaire, le collectif *Propublica* rapporte que l'outil de prévention de récidivisme *COMPAS*, utilisé en justice prédictive, présente un taux de faux positifs dans la population afro-américaine significativement plus grand que dans la population caucasienne (Angwin, Larson, Mattu et Kirchner, 2016).

La sous-représentation de groupes dans les données d'entraînement peut aussi conduire à des biais non-désirés, comme ce fut notamment le cas de l'outil de recrutement d'Amazon qui était discriminant envers la gent féminine parce que cette catégorie était sous-représentée dans les données d'entraînement du modèle (Dastin, 2018)). Pareillement, son outil de reconnaissance faciale avait associé plusieurs membres du congrès américain à des criminels (Brandom, 2018). On peut aussi citer les standards de beauté souffrant d'un manque criant de diversité et de caractéristiques fa-

ciales (Muraski, 2021)), et tel que démontré par la recherche *Gender Shades* (Buolamwini et Gebru, 2018).

De même, puisque l'apprentissage automatique libère la personne qui la conçoit le programme de la programmation explicite, il peut aussi conduire à l'apprentissage de modèle de type "boîte noire" qui effectuerait des prédictions sans que la personne utilisatrice ne soit en mesure de comprendre le raisonnement derrière la prise de décision. Beaucoup d'autres exemples (Vincent, 2020 ; Yudkowsky et al., 2008) mettent en avant la capacité de la machine automatique à utiliser des corrélations existantes, qui sont non pertinentes et non souhaitables pour la tâche à résoudre. Dans ces exemples, la machine semblait être performante en termes de qualité de prédiction durant sa conception, mais avait des performances médiocres en phase de déploiement. Cette différence de performance a permis de révéler l'utilisation de ces corrélations non néfastes pour la tâche et difficilement détectables à cause de l'opacité des modèles appris.

La transparence et l'explicabilité des prédictions du modèle appris sont ainsi primordiales, non seulement pour s'assurer de la pertinence des corrélations apprises, mais aussi afin de protéger la personne utilisatrice dans les décisions prises ou qu'elle aura à prendre. Plusieurs cadres légaux ont été mis en place ou sont en cours de préparation afin d'encadrer cette technologie et d'en limiter les dérives. On peut citer par exemple le projet de loi américain «*Algorithmic Accountability Act of 2022*» (Wyden, 2022), qui propose un encadrement de l'utilisation des outils de prise de décision automatique, et requiert une évaluation d'impact de l'utilisation de ces outils par les entreprises qui les développe ou en font usage.

En Europe, le «Règlement Général sur la Protection des Données (RGPD)» (Parlement Européen, 2018) entré en vigueur en mai 2018 souligne aussi les aspects de contrôle et de responsabilité de l'apprentissage machine. Enfin, plus récemment, la *Proposition de Législation sur l'Intelligence Artificielle (PLIA)* (Parlement Européen, 2021) va plus loin dans l'encadrement de l'intelligence artificielle. On retrouve par exemple dans ce projet de législation une classification des applications d'apprentissage automatique et des systèmes de décision automatique, avec pour chaque niveau de classification, des suggestions pour la gestion de risques et les obligations entourant l'usage du système.

Au Canada et au Québec, de nouvelles réglementations sur la protection des renseignements personnels et sur la réglementation des systèmes automatiques entrent aussi en vigueur : le projet de loi

C-11 (Ministre de l'innovation des sciences et de l'industrie, 2020), la directive sur les décisions automatisées [1] promue par le gouvernement fédéral et la loi 64 promue par le gouvernement du Québec. Ces cadres légaux établissent un cadre formel dans lequel l'utilisation de l'apprentissage automatique ne doit pas se faire au détriment de la dignité de la personne. Ainsi, les décisions prises au sujet d'une personne doivent pouvoir être justifiées afin de prévenir les dérives d'une utilisation abusive. Concrètement, la loi 64 propose un cadre dans lequel les personnes doivent être informés des décisions prises à leur sujet et leur offre la possibilité d'exprimer leur opinion à propos de la décision et du système automatique. Concernant le projet de loi C-11 (qui propose des amendements similaires au RGPD et à la PLIA), le gouvernement fédéral canadien a crée un outil [2] pour assister les entreprises dans l'évaluation de l'impact de leurs algorithmes sur les personnes humaines et la société.

Toutes ces législations s'appuient notamment sur d'autres préexistantes, telles que les textes définissant les droits fondamentaux ou encore les lois anti-discrimination et étendent ces cadres légaux dans le contexte d'outils automatiques capables de prendre des décisions sans intervention humaine. Ainsi, on retrouve par exemple dans la *Loi canadienne sur les droits de la personne (L.R.C. (1985), ch. H-6)* (Justice, 1985) [3] : «*les motifs de distinction illicite sont ceux qui sont fondés sur la race, l'origine nationale ou ethnique, la couleur, la religion, l'âge, le sexe, l'orientation sexuelle, l'identité ou l'expression de genre, l'état matrimonial, la situation de famille, les caractéristiques génétiques, l'état de personne graciée ou la déficience.* ». En d'autres termes, les outils automatiques ne peuvent pas être utilisés pour outrepasser les droits fondamentaux des personnes.

Ces droits fondamentaux incluent aussi le droit au respect de la vie privée (article 12 de la Déclaration universelle des droits de l'homme (Nations Unies, 2008)), qui peut être sévèrement menacé par les systèmes d'apprentissage automatique. En effet, ces systèmes nécessitent de considérables masses de données pour être fonctionnels. La collecte de ces données peut se faire de manière insidieuse, sans le consentement de la personne concernée, et même dans le cas où cette collecte est faite de manière légale, les inférences (profilage ou « *behavioural profiling* ») qui sont faites par l'association de données collectées séparément, mais qui sont ensuite agrégées dans un profil commun, ou la

1. `https://www.tbs-sct.canada.ca/pol/doc-fra.aspx?id=32592`

2. `https://open.canada.ca/aia-eia-js/?lang=en`

3. `https://laws-lois.justice.gc.ca/fra/lois/h-6/TexteComplet.html`

personnalisation intersystèmes peut dangereusement porter atteinte à la vie privée (Toch, Wang et Cranor, 2012).

On assiste ainsi à plusieurs dérives telles que l'ingérence dans les élections (Judge et Pal, 2021 ; Lys, 2019), la surveillance par des acteurs étatiques ou encore par des entreprises (Drinhausen et Brussee, 2021 ; Fung, 2017), les campagnes de marketing ciblées (Dixon, 2013 ; privacyinternational, 2019) ainsi que les pratiques peu respectueuses de la vie privée des compagnies d'assurance (Gent, 2017 ; Leefeldt et Danise, 2021 ; Smith, 2021). Le respect de la vie privée apparaît ainsi comme étant aussi fondamental pour la construction de systèmes de décision automatique socialement acceptables (Barker, Askari, Banerjee, Ghazinour et al., 2009 ; Introna, 1997 ; Perera, Ranjan, Wang, Khan et al., 2015 ; Rachels, 1975), comme le soulignent les articles des diverses régulations sur l'apprentissage automatique susmentionnées (Parlement Européen, 2018, 2021) ou encore les principes du *U.S. Privacy Act (1974)*.

En plus des cadres légaux mis en place par les gouvernements, les entreprises, elles aussi, lancent plusieurs initiatives portant sur la régulation des systèmes de décision automatique. De manière générale, on retrouve pour la plupart de ces initiatives, les principes directeurs suivants qui orientent la recherche et l'utilisation des systèmes algorithmiques : l'intervention humaine, la robustesse, la sécurité, le respect de la vie privée, la transparence, la responsabilité, la non-discrimination ainsi que le bien-être sociétal et environnemental (Facebook, 2022 ; Google, 2022 ; Microsoft, 2022).

Enfin, on observe depuis quelques années la montée importante d'un nouveau champ de recherche appelé *éthique de l'intelligence artificielle*, regroupant des scientifiques provenant de domaines différents, tels que l'informatique, l'éthique, le droit et la sociologie. Notre thèse s'inscrit dans le cadre de ce domaine de recherche émergent. Plus précisément, nous nous intéressons à la problématique de protection des renseignements personnels, tout en assurant l'équité et la non-discrimination lors de l'utilisation des systèmes utilisant l'apprentissage automatique.

Le premier chapitre de notre thèse consistera en une revue de littérature des techniques d'amélioration de l'équité dans la prise de décision par des techniques d'apprentissage automatique. Plus précisément, nous commencerons par présenter brièvement l'historicité de cette problématique, les principales sources de discrimination au sein des données, les différentes définitions d'équité et leurs formulations mathématiques. Ensuite, nous dresserons un état de l'art des méthodes permettant d'améliorer l'équité et de prévenir la discrimination.

Le deuxième chapitre de cette thèse détaillera notre première contribution «Local Data Debiasing for Fairness Based on Generative Adversarial Training (Aïvodji, Bidet, Gambs, Ngueveu et al., 2021)» publiée dans le numéro spécial *Interpretability, Accountability and Robustness in Machine Learning* du journal *Algorithms (2021)*. Notre approche, appelée GANSan, consiste en un prétraitement des données pour réduire les discriminations (mesurées à partir un attribut spécifique) dans un ensemble de données avant son utilisation par des modèles à entraîner. Nous avons pu montrer l'efficacité de notre solution sur plusieurs cas d'utilisation. Ensuite, nous avons évalué la perte de performance en termes de qualité de prédiction si notre outil opère sur une portion limitée de l'ensemble de données.

Le troisième chapitre consistera en l'application de GANSan pour assurer la protection d'un renseignement personnel dans le contexte des données issues de capteurs : «DYSAN : Dynamically sanitizing motion sensor data against sensitive inferences through adversarial networks (Boutet, Frindel, Gambs, Jourdan et al., 2021)». Cette approche permet d'obtenir de meilleures performances que celles obtenues par des méthodes similaires, grâce à la sélection dynamique du meilleur modèle pour prévenir l'inférence de l'information sensible dans les sous-ensembles de données de chaque personne utilisatrice de l'approche. Notre solution affecte peu la durée de vie du terminal mobile sur lequel elle est déployée, permet la protection des données contre l'inférence d'attributs et maintient un haut niveau d'utilité des données décorrélées du renseignement personnel choisi (ex : le genre). Ce travail de recherche a été publié dans la conférence *ACM Asia Conference on Computer and Communications Security (AsiaCCS)* en 2021.

Le dernier chapitre de notre thèse portera sur un travail actuellement en cours de soumission *Fair-Mapping* (FM), qui est une technique d'amélioration de l'équité par le transport optimal d'une distribution choisie vers une autre de l'ensemble de données, par exemple la distribution du groupe recevant le plus d'avantages sociaux. Cette approche, en plus d'attribuer à toutes les personnes de l'ensemble de données des caractéristiques voulues, préserve l'aspect réaliste des données en s'assurant que le résultat de la procédure de transport correspond à une distribution réelle. Ce travail est actuellement en cours de révision pour une soumission dans la revue spécialisée *Safe and Fair Machine Learning* du journal *Machine Learning*.

Enfin, nous terminerons cette thèse avec un point de vue critique sur l'ensemble de nos réalisations. Nous discuterons de certaines limites que nous avons pu observer dans les méthodes d'amélioration de l'équité, et présenterons quelques pistes de recherches futures.

CHAPITRE I

L'ÉQUITÉ DE LA PRISE DE DÉCISION AUTOMATIQUE

L'équité, par définition, est le caractère de ce qui est fait avec justice et impartialité [1]. Par exemple, des décisions prises entre des personnes sont équitables lorsque ces décisions ont été prises (1) de manière juste entre ces personnes, c'est-à-dire en attribuant à chacun la décision qui lui revient en fonction de son droit et son mérite, et (2) de manière impartiale, c'est-à-dire que ces décisions ont été rendues sans *à priori* ou préjugés, sans favoritisme et en s'appuyant uniquement sur les critères justes pour la prise de ces décisions.

Dans le cadre de l'apprentissage automatique, l'équité de la prise de décision se traduit par l'impartialité ou l'invariance de cette prise de décision par rapport à certains facteurs jugés sensibles. Ces facteurs peuvent correspondre aux attributs ou caractéristiques personnelles sur la base desquels la prise de décision est interdite par la loi (par exemple, la *croyance religieuse*). Toutefois, arriver à obtenir un processus de décision automatique impartial peut être très difficile. En effet, la prise de décision automatique peut être affectée à différents niveaux par des distorsions ou biais qui peuvent être, dans certains cas, difficilement décelables. Ainsi, la recherche d'un système capable de prendre des décisions équitables automatiquement reste un domaine de recherche très actif.

Hutchinson et Mitchell (2019) ont montré que ce domaine de recherche a été exploré depuis les années 1950, mais avec des variations importantes dans l'intérêt qui y a été porté. Ils constatent en particulier que plusieurs des problématiques étudiées de nos jours existaient déjà dans les années 1960, bien que les biais étaient majoritairement observés au sein des domaines de l'éducation et de l'emploi. Des études avaient été menées pour aborder les problématiques d'équité, leur définition et leur quantification dans ces contextes. Après une baisse de l'intérêt porté à ce domaine de recherche

1. https://www.larousse.fr/dictionnaires/francais/équité/30712

dans les années 1970, notamment à cause de l'incapacité à identifier sans ambigüité les cas de discrimination, l'absence de procédure pour éviter la discrimination et les opinions divergentes sur les mesures d'équité, les questions liées à ce domaine ont été relancé par le débat public dans les années 1980 par des études controversées portant l'influence de l'origine ethnique sur l'intelligence, notamment dans la recherche sur les biais en test mentaux (Jensen, 1980) (établissant des liens entre la génétique et ces tests). De nos jours, le nombre de publications et de cas d'études liés aux enjeux éthiques en apprentissage machine, augmente de manière importante. Plusieurs conférences et journaux sur ces enjeux ont vu le jour, ainsi que des principes directeurs de la recherche en intelligence artificielle, mis en avant par des laboratoires spécialisés dans le domaine.

L'objectif de ce premier chapitre sera donc d'apporter une vision d'ensemble dans ce domaine de recherche. Nous commencerons par introduire les notations et définitions de bases utilisées en apprentissage automatique, qui nous seront aussi utiles tout au long cette thèse. Ensuite, nous définirons les notions et les concepts nécessaires à la compréhension des recherches menées dans la prise de décision équitable par des systèmes automatiques. Le reste du chapitre concernera les mécanismes par lesquels les biais ou distorsions pourraient être introduits durant le processus de prise de décision, ainsi que les moyens et méthodes pour lutter contre ces biais. En particulier, nous y présenterons les types de biais et leurs origines potentielles (partie 1.3), les différentes formes que peuvent prendre l'équité des décisions ainsi que les métriques pour la quantifier (partie 1.4), les ensembles de données les plus utilisés pour prouver l'efficacité des méthodes (partie 1.5), et enfin les méthodes et les algorithmes pour améliorer l'équité dans la prise de décision (partie 1.6), ainsi que leur relation avec les méthodes de génération des données (partie 1.7) et de protection des renseignements personnels (partie 1.8).

1.1 Symboles et notations

Nous considérons dans cette thèse des données structurées, définies par un ensemble de données R composé des profils de N différentes personnes. Chaque profil r_i ($i \in \{1, \ldots, N\}$) est décrit par un ensemble de d attributs. L'ensemble des attributs est composé :

— D'un ou plusieurs (de 2 à k) *attributs sensibles* noté(s) S_j ($j \in \{1, \ldots, k\}$). Les attributs sensibles sont les attributs à partir desquels la discrimination est formulée, mesurée et évaluée. Par exemple, il s'agit d'attributs tels que l'*origine ethnique*, le *sexe* ou encore l'*âge*. Dans notre contexte, nous considérons les attributs sensibles comme étant binaires ou multivalués.

Parmi l'ensemble des groupes définis par les valeurs de l'attribut sensible (cas d'un seul attribut sensible binaire) ou par les combinaisons de valeurs de ces attributs (cas de plusieurs attributs sensibles), nous identifierons le groupe privilégié comme étant le groupe ayant des propriétés jugées les plus utiles pour les besoins de la personne ou entité utilisatrice des données, parmi tous les groupes. Le rationnel du choix du groupe peut être basés sur les avantages reçus ou discrimination subie par les membres de ce groupe comparativement aux autres, sur la taille de ce groupe par rapport aux autres, ou autre. Nous simplifierons les notations en considérant que l'ensemble de donnée ne contient qu'un seul attribut sensible S pouvant prendre k différentes valeurs. Ainsi, dans le cas où on ne considère qu'un seul attribut sensible binaire, S prendra les valeurs s_1 et s_0. Dans le cas où l'attribut sensible est multivalué (ou si l'on considère plusieurs attributs sensibles), S représentera l'ensemble des combinaisons de valeurs des attributs sensibles, et prendra les valeurs s_i, avec s_i correspondant à une combinaison entre ces attributs. Nous identifierons le groupe privilégié avec la valeur s_1, tandis que chacun des autres groupes sera identifié par la valeur s_i, $i \neq 1$. Ces autres groupes seront considérés comme des groupes protégés.

— D'un attribut de décision binaire dénoté Y. Cet attribut pourrait par exemple correspondre à la décision (*positive* ou *négative*) qu'un(e) recruteur(euse) assignerait à un candidat à l'issue de l'évaluation du dossier de candidature. Elle représente aussi la décision donnée par un système automatique qui remplacerait la personne responsable du recrutement dans l'exemple précédent. On distinguera ainsi les décisions positives $Y = 1$ de celles négatives $Y = 0$. L'espace de l'ensemble des décisions est symbolisé par $\mathcal{Y}$.

— L'ensemble des autres attributs qui ne sont ni sensibles, ni de décision, sera noté A. Cependant, ces attributs peuvent être corrélés aux attributs sensibles et à l'attribut de décision. En effet, c'est à partir de ces attributs que la décision Y est inférée par le système de décision automatique. De même, il peut être possible d'inférer les valeurs de l'attribut S à partir de A avec des meilleurs résultats que ceux obtenus par un modèle de référence qui prédirait de manière aléatoire ces attributs sensibles. On désignera par $\mathcal{A}$ l'espace de ces attributs, c'est-à-dire l'univers contenant tous les attributs non sensibles et leurs valeurs, sur lequel on peut définir une ou plusieurs métriques de distances notées $d_{\mathcal{A}}$.

Notons que nous désignerons par les lettres majuscules A, Y et S les variables aléatoires représentant les attributs, tandis que les lettres en minuscule a, y et s représenterons (dans le même ordre) la réalisation de ces variables, c'est-à-dire des valeurs de ces attributs. La liste des abréviations

(présentée dans les premières pages de cette thèse), des sigles et des acronymes résume les notations utilisées dans le reste de cette thèse.

1.2 Apprentissage supervisé

L'apprentissage supervisé, avec l'apprentissage non supervisé et l'apprentissage par renforcement, forment les trois grandes formes d'apprentissage automatique (Jordan et Mitchell, 2015). Nos travaux concernent uniquement l'apprentissage supervisé, qui a pour but de modéliser la distribution conditionnelle $P(Y|A)$. En d'autres termes, l'apprentissage supervisé consiste en la construction d'un modèle h mettant en relation les intrants et les extrants donnés sous forme de paires (a_i, y_i). Par exemple, le modèle peut être construit pour prédire le niveau de revenus d'une personne (Y) à partir des attributs (A) tels que le niveau d'éducation, le type d'emploie occupé, le nombre d'années d'expériences. Le sens de modèle en apprentissage supervisé peut être interprété comme une forme de représentation des données permettant décrire de manière précise les relations entre les extrants y_i et les intrants a_i. Cette représentation peut prendre la forme d'une série d'équations qui permettent d'exprimer y_i en fonction de a_i. L'apprentissage supervisé diffère principalement des deux autres par le fait que les extrants utilisés (les sorties souhaitables du modèle) durant la phase de construction du modèle sont connus et mesurables. Plusieurs types de fonction peuvent être appris. Par exemple, on distingue les arbres de décisions (Quinlan, 1986), les réseaux de neurones (Cortes et Vapnik, 1995) ou les machines à vecteurs de support (Wang, 2005). Ainsi, on peut considérer un modèle de prédiction ou classifieur h, qui reçoit en entrée des données A et prédit la décision Y elle-même, ou produit en sortie un score correspondant à la probabilité d'obtenir une décision positive ($h(a_i) = P(y_i = 1|A = a_i)$).

À partir des vraies valeurs de la décision Y et des prédictions réalisées par le modèle $\hat{Y}$, on peut définir la matrice de confusion (tableau 1.1) qui résume les performances du modèle. À partir de celle-ci, on peut définir plusieurs métriques, telles que l'exactitude de prédiction (définition 1), la précision (définition 2) ainsi que les taux de vrais positifs et les taux de faux négatifs (définition 3).

Soit un classifieur h entraîné à prédire un attribut Y à partir des attributs A dans l'ensemble de données R :

Définition 1 (Exactitude de prédiction du classifieur)**.** *L'exactitude de prédiction de h à prédire Y*

TABLEAU 1.1 – Matrice de confusion de la prédiction de l'attribut binaire Y par le modèle h. Le symbole "#" devant les notations telles que "vrais positifs" indique le nombre d'éléments de l'ensemble représenté par la notation (par exemple, le nombre de vrais positifs)

		$\hat{Y}$		Total
		$\hat{y}_0$	$\hat{y}_1$	
Y	y_0	#vrais négatifs	*#faux positifs*	n_{y_0}
	y_1	*#faux négatifs*	*#vrais positifs*	n_{y_1}
Total		$n_{\hat{y_0}}$	$n_{\hat{y_1}}$	N

se définit comme :

$$\begin{aligned} Acc_Y(h,Y) &= \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}(y_j = h(a_j)) \\ &= \frac{\#vrais\ positifs + \#vrais\ négatifs}{N} \end{aligned} \tag{1.1}$$

où y_j représente la valeur de l'attribut Y du profil r_j, a_j les valeurs des attributs A et $\mathbb{1}(x = y))$ une fonction indicative qui prend la valeur 1 si x est égal y et 0 sinon. L'exactitude de prédiction pourra aussi être noté par le sigle Acc.

Définition 2 (Précision du classifieur)**.** *La précision d'un classifieur est définie comme la proportion de prédictions positives du classifieur qui sont correctes.*

$$\begin{aligned} Precision(h,Y) &= P(Y = 1|\hat{Y} = 1) \\ &= \frac{1}{\#(\hat{Y} = 1)} \sum_{j=1|h(a_j)=1}^{\#(\hat{Y}=1)} \mathbb{1}(y_j = 1) \\ &= \frac{\#vrais\ positifs}{\#vrais\ positifs\ +\ \#faux\ positifs} \end{aligned} \tag{1.2}$$

où $\#(\hat{Y} = 1)$ représente le nombre d'éléments ayant une décision prédite positive et $\hat{Y}$ représente la prédiction de Y par h : $\hat{Y} = h(A)$.

Définition 3 (Taux de vrais positifs et taux de faux négatifs)**.** *Le taux de vrais positifs (Vp ou encore $Tp.Rate$) de h est défini comme la proportion de décisions positives qui sont correctement*

prédites :

$$\begin{aligned} Vp(h,Y) &= P(h(a_j)=1|Y=1) \\ &= \frac{\sum_{j=1}^{\#(Y=1)} \mathbb{1}(1=h(a_j))}{\#(Y=1)} \\ &= \frac{\#\textit{vrais positifs}}{\#\textit{vrais positifs} + \#\textit{faux négatifs}} \end{aligned} \tag{1.3}$$

avec $\#(Y=1)$ *définissant la cardinalité de l'ensemble des profils ayant des décisions positives dans l'ensemble de données.*

Le taux de faux positifs (Fp ou encore Fp.Rate) de f *correspond à la fraction de décisions négatives qui sont prédites positives :*

$$\begin{aligned} Fp(h,Y) &= P(h(a_j)=1/Y=0) \\ &= \frac{\sum_{j=1}^{\#(Y=0)} \mathbb{1}(1=h(a_j))}{\#(Y=0)} \\ &= \frac{\#\textit{faux positifs}}{\#\textit{faux positifs} + \#\textit{vrais négatifs}} \end{aligned} \tag{1.4}$$

Typiquement, l'apprentissage supervisé nécessite la division de l'ensemble de données D en trois sous-ensembles : un ensemble d'entraînement, un ensemble de validation et un ensemble de test (Géron, 2019).

- L'ensemble d'entraînement est utilisé pour construire le modèle. Par exemple, à partir de cet ensemble, un algorithme d'apprentissage de type arbre de décision construit les branches et les noeuds sur lesquels la division des données doit être faite. Les réseaux de neurones y modifient notamment leurs paramètres internes pour arriver à prédire correctement la décision associée aux différents profils de cet ensemble d'entraînement.
- L'entraînement des modèles nécessite plusieurs hyperparamètres qui définissent les choix et les hypothèses nécessaires pour la conception de ce modèle qui sont choisis grâce à l'ensemble de validation. Ces hyperparamètres correspondent par exemple à la profondeur maximale de l'arbre de décision souhaitée, le nombre de couches du réseau de neurones ou encore le nombre de noeuds de chaque couche. Ainsi, plusieurs modèles sont construits sur l'ensemble d'entraînement et l'ensemble de validation permet de choisir le modèle et donc la combinaison d'hyperparamètres permettant d'obtenir les meilleurs résultats suivant des métriques de performances définies précédemment.
- Le meilleur modèle étant choisi sur l'ensemble de validation, la capacité de généralisation du modèle, c'est-à-dire sa capacité à être performant sur des données qui n'ont jamais été

utilisées, ne peut pas être mesuré directement sur cet ensemble. Ainsi l'ensemble de test permet ainsi de mesurer la capacité de généralisation du modèle afin de s'assurer que celui-ci est en mesure d'être performant une fois déployé.

Il existe d'autres procédures de mesures de la qualité du modèle telles que la validation croisée ou k-validation qui permettent d'évaluer les performances du modèle dans une configuration plus complexe (Géron, 2019). Nous introduirons cette procédure dans le chapitre 2, partie 2.2.4.

Ci-après, nous allons définir des mécanismes par lesquels la discrimination peut exister et être propagée ainsi que des métriques permettant de quantifier l'équité.

1.3 Origine des biais

La discrimination est définie comme le fait de distinguer et de traiter différemment (le plus souvent de manière défavorable) quelqu'un ou un groupe par rapport au reste de la collectivité ou par rapport à une autre personne [2]. Celle-ci peut-être introduite dans le cadre de l'apprentissage automatique par l'intermédiaire de biais, conduisant par conséquent à des résultats différents en fonction de l'affinité ou orientation de ces biais face à des personnes (ou des données) de différents groupes (Hutchinson et Mitchell, 2019 ; Romei et Ruggieri, 2014). Comme définis précédemment, les biais sont des distorsions qui affectent le processus de décision et l'empêche d'être impartial. Les biais qui se retrouvent au sein des modèles entraînés peuvent provenir de plusieurs sources, ils peuvent se complémenter pour amplifier le biais final, ou encore s'annuler, laissant croire à la non-existence de biais.

Plus précisément, les biais peuvent exister à plusieurs niveaux dans les différentes étapes du pipeline de l'apprentissage automatique. On peut distinguer les biais durant la phase de spécification du problème, les biais dans les données collectées, ceux de l'étape de modélisation, d'évaluation et de validation des données et finalement les biais qui peuvent être introduit durant la phase de déploiement (Fazelpour et Danks, 2021).

1.3.1 Biais dus à la spécification du problème

L'apprentissage automatique est souvent au coeur de la construction de systèmes de décision automatique servant à la résolution de problèmes. Cependant, certains objectifs sont très complexes et

2. https://www.larousse.fr/dictionnaires/francais/discrimination/25877

difficilement exprimables de façon concrète et tangible. Aussi, les attributs utilisés pour la réalisation de la tâche peuvent parfois capturer de manière très limitée le but à atteindre. Par exemple, dans le contexte de l'éducation, la réussite scolaire est un objectif difficilement mesurable. En effet, cet objectif peut être composé de plusieurs aspects variés et subjectifs (par exemple, la cote R au Québec, le GPA - moyenne pondérée des résultats scolaires, les activités parascolaires, etc.). Ainsi, d'un centre scolaire à l'autre, les mesures d'évaluation de la réussite scolaire pourraient être très différentes. Ces mesures pourraient ainsi être des sources de biais, étant donné que les personnes qui étudient dans un centre peuvent ne pas avoir de bonnes performances suivant les mesures utilisées par le centre en question, mais en auraient eu de très bonnes suivant d'autres mesures utilisées dans d'autres centres. Le caractère subjectif de la mesure introduit ainsi des perspectives différentes ou distorsions dans l'évaluation de l'objectif. Dans la spécification du problème, on distingue comme source potentielle de biais :

— *La sélection des critères.* Pour une même finalité, les critères choisis et utilisés pour faire la prédiction peuvent conduire à des décisions différentes. En effet, Roselli, Matthews et Talagala (2019) illustrent cette situation dans le cadre de la prédiction de la réussite des élèves dans un cadre scolaire. Deux combinaisons différentes de critères ne mèneraient pas systématiquement à des résultats identiques, à cause de la variété des métriques utilisées (résultats d'examen, classement, lettres de recommandation, GPA), des observations (telles que des remarques personnelles) qui peuvent être difficilement quantifiables ou encore de l'ignorance de l'existence de certaines métriques mieux adaptées pour la tâche.

— *Les données substituts.* De manière générale, les données substituts représentent l'ensemble des attributs auxquels les personnes qui conçoivent ou développement les modèles ont accès. Friedler, Scheidegger et Venkatasubramanian (2016) caractérisent ces données par «espace observable» et considèrent qu'elles constituent des moyens d'accès indirect à la propriété qu'on cherche à mesurer. Par exemple, dans le système nord-américain, il est courant d'utiliser les scores de crédit pour évaluer la santé financière d'une personne. La limitation à ce score constituerait une perte d'information dans la mesure de la propriété *santé financière*, car elle ne représente qu'un aspect très partiel des situations personnelles et sociales (et sociétales) auxquelles une personne est confronté.

— *L'espace de décision.* Comme mentionné précédemment, l'espace de décision fait référence à l'objectif à atteindre. Or la propriété finale qu'on cherche à mesurer n'est pas forcément exprimable quantitativement ou encore les modèles mathématiques ne seront pas forcément

en mesure de la capturer. Par conséquence, des données substituts peuvent-être utilisées pour mesurer cette propriété, ce qui engendre des problèmes lorsque les prédictions faites à partir des données substituts sont utilisées comme conclusions sur la propriété.

1.3.2 Biais dans les données d'entraînement

Les données d'entraînement peuvent aussi comporter des biais. Ceux-ci peuvent provenir du processus de collecte des données, être du à la précision des nombres en virgule flottante [3] ou encore aux préjugés des personnes collectant ou partageant leurs informations. L'étude de Mehrabi, Morstatter, Saxena, Lerman et al. (2021) fait un tour d'horizon des possibles origines de biais dans l'apprentissage automatique. Ils en identifient 25 parmi lesquelles 13 sont directement liés aux données utilisées pour l'apprentissage. Ces biais peuvent être classifiés en différentes sous-catégories, notamment les biais historiques, les biais de manipulation, ceux de non-représentativité et enfin les biais d'analyse.

1. *Les biais historiques.* Ce type de biais trouve son origine dans les problèmes sociaux et sociétaux qui existent dans le monde réel, où le traitement réservé à certains groupes de la population est significativement différent de celle-ci. Ces différences de traitements créent des inégalités entre les groupes de la population et peuvent se perpétuer dans le temps. Par exemple, la discrimination à laquelle ont fait face les groupes de personnes de couleurs aux États-Unis est à l'origine de leur manque d'accès à certaines ressources, et ainsi à leur sous-représentation dans certains groupes sociaux. Les données peuvent en conséquence capturer ces biais, de même que la personne qui opère les systèmes peut être sujette à des biais s'exprimant sous forme de stéréotypes.
2. *Les biais de manipulation.* Ce type de biais comprend aussi bien les transformations fallacieuses des données que les erreurs d'échantillonnage ou d'enregistrement de données. Les biais de manipulation apparaissent par exemple au travers des biais de mesures, qui, selon Mehrabi, Morstatter, Saxena, Lerman et al. (2021), correspondent à la manière dont on choisit, utilise et mesure des caractéristiques particulières. Ainsi, le choix des attributs de décision peut être une simplifications excessives de constructions plus complexes comme par exemple le fait d'utiliser la variable GPA comme seule mesure de potentiel d'un(e)

3. Les opérations sur les nombres réels peuvent conduire à des résultats difficiles à représenter en mémoire des systèmes informatique, et des arrondis peuvent aussi être faits, ce qui pourraient fausser les résultats finaux

étudiant(e). Suresh et Guttag (2019) généralise ces biais de mesures à la variation de la méthode de mesure et à la variation de la précision des mesures entre les groupes.

3. *Les biais de non-représentativité.* Ces biais émergent principalement dans la situation où l'échantillon choisi ne représente pas la population dans son ensemble, conduisant ainsi à une sous-estimation ou à une surreprésentation (Silva et Kenney, 2018 ; Suresh et Guttag, 2019) des groupes. En pratique, les biais de non-représentativité peuvent tirer leur source de plusieurs autres biais tels que le biais d'échantillonnage, les biais sociaux qui influencent les choix de populations ou leurs proportions, les biais d'autosélection, les biais de liaison, ainsi que les biais de production de contenu (Mehrabi, Morstatter, Saxena, Lerman et al., 2021).

4. *Les biais d'analyse.* Il s'agit de biais provenant des analyses effectuées sur l'ensemble de données lors de son étude. Dans ce cas, pour un même ensemble de données, plusieurs conclusions différentes et parfois opposées peuvent être tirées selon l'angle d'observation choisi. Par exemple, en observant les données d'admission dans un collège on pourrait observer que le taux d'admission de personnes originaires d'un pays A est très inférieur par rapport à celui d'autres groupes. Cependant, à l'échelle de chaque faculté dans le collège, on peut observer que le taux d'admission des personnes de ce pays A est le plus élevé. Ainsi, selon l'échelle choisie, on peut arriver à des conclusions très différentes. Les biais d'analyse peuvent trouver leur origine dans le paradoxe de Simpson (Blyth, 1972) [4], le sophisme des données longitudinales [5], les biais de comportement (comportements différents d'une personne en fonction de la plateforme sur laquelle il se trouve (Mehrabi, Morstatter, Saxena, Lerman et al., 2021)) ainsi que dans les biais d'omission de critère.

1.3.3 Biais dans la modélisation et validation des résultats

Les données ne sont pas les seuls éléments dans lesquels les biais peuvent être introduits. En effet, en supposant les données exemptes de tout biais non désirable, les hypothèses utilisées dans la concep-

4. Le paradoxe de Simpson est un phénomène statistique où une association entre deux variables dans une population émerge, disparaît ou s'inverse lorsque la population est divisée en sous-groupes de manière différente.

5. Le sophisme des données longitudinales est le fait de considérer des données transversales (échantillon pris à un instant spécifique dans le temps), comme si elles étaient longitudinales (données collectées à plusieurs moments sur une période étendue)

tion de l'algorithme d'entraînement de modèles peuvent introduire des biais dans les prédictions réalisées par les modèles construits par ces algorithmes. Ces hypothèses peuvent refléter les objectifs des parties prenantes, qui contribuent aussi à l'orientation des décisions à prendre. De même, le modèle entraîné peut être sujet à des erreurs de fonctionnement (Vincent, 2018, 2020), sans tenir compte de la subjectivité des personnes chargées d'évaluer le modèle, qui peuvent parfois être en conflit d'intérêts. En effet, une entreprise proposant un nouveau produit chercherait à montrer les avantages de celui-ci dans différentes situations, sans avoir suffisamment de recul pour en voir les limites. Les biais de la phase de modélisation peuvent être regroupés en trois grandes catégories : *les biais d'agrégation*, les *biais d'attention algorithmique* et les *biais de traitement algorithmique*.

Suresh et Guttag (2021) considèrent que les biais d'agrégation apparaissent lorsqu'un modèle unique est utilisé pour le traitement de tous les groupes sensibles présents dans les données. L'hypothèse sous-jacente étant que l'association entre les intrants et les prédictions soient les mêmes pour tous les groupes. L'exemple utilisé pour illustrer ce type de biais est notamment la détection de diabète, qui s'appuie largement sur une caractéristique précise. Cependant, cette caractéristique varie significativement entre les groupes et le genre. Le biais d'agrégation consisterait ainsi à considérer les valeurs de cette caractéristique comme ayant une signification unique et identique pour tous. Ce problème peut être illustré par le fait que ces modèles entraînés à partir d'une erreur moyenne ont parfois des performances médiocres sur les données considérées comme anomalies (*outliers*). En effet, ces anomalies sont différentes des autres données de l'ensemble d'entraînement, ce qui conduit au fait que l'erreur calculée sur ces données est élevée. Toutefois, étant donné que les erreurs calculées sur les autres données de l'ensemble d'entraînement sont très petites par rapport aux erreurs calculées sur les anomalies, l'impact de ces erreurs est caché dans la moyenne.

Concernant les biais d'attention, ils apparaissent à la suite de l'utilisation (ou au contraire de l'absence d'utilisation) de certains attributs dans le système de décision automatique. Par exemple, Danks et London (2017) considèrent la situation dans laquelle des attributs peuvent être corrélés à des attributs sensibles, et par conséquent biaiser un modèle qui ne l'aurait été sans l'utilisation de ceux-ci. Concrètement, dans le cadre de la justice prédictive, bien que des attributs démographiques tel que le genre ou l'origine ethnique, soient collectés, les utiliser à des fins de prédiction pourrait résulter en des risques de discrimination. Ainsi, des personnes chargées du développement d'un modèle prédictif pourraient choisir d'ignorer ces attributs.

Enfin, les biais de traitement algorithmique résultent des critères formels et/ou mathématiques uti-

lisés pour permettre à la machine de réaliser la tâche souhaitée. Concrètement, il s'agit de jugements de valeur apportés sur le choix des métriques à optimiser, les configurations à choisir, mais surtout les choix de conception effectués, par exemple, les compromis à faire entre plusieurs critères (Danks et London, 2017 ; Fazelpour et Danks, 2021). En effet, les personnes chargées du développement de modèles doivent souvent trouver un équilibre entre la qualité du modèle, le temps de conception et d'exécution de celui-ci ainsi que les ressources limitées. Ainsi, gagner quelques points de performance en termes de qualité de prédiction peut s'avérer très coûteux en temps et ressources, mais cette amélioration de performance pourrait conduire à un meilleur traitement de certaines personnes. Ainsi dans le domaine de l'équité, des compromis sont aussi faits entre la qualité de prédiction du modèle et l'équité globale du système. Par exemple, plusieurs modèles peuvent-être obtenus pour la prédiction de la qualité de crédit des personnes : un premier modèle ayant une exactitude de prédiction jugée excellente, mais dont les prédictions sont fortement discriminantes envers les personnes de moins de 30 ans, un second modèle dont l'exactitude de prédiction est de 10% inférieure au premier mais dont la qualité de prédiction reste la même pour les personnes de moins de 30 et celles plus âgées. Le choix du modèle à utiliser reflétera les objectifs de décision, et par ricochet ses biais de traitement algorithmique.

Les *biais de validation ou biais d'évaluation* introduisent de la subjectivité dans le modèle à entraîner, qui peut être due aux besoins métier et aux décisions des parties prenantes. Pour Suresh et Guttag (2019), les biais d'évaluation proviennent de l'inadéquation entre les données d'entraînement du modèle et la population cible. Dans notre recherche, nous considérons que les biais d'évaluation peuvent trouver leur source dans le large éventail des biais d'interprétations des personnes chargée de l'évaluation du modèle entraîné (Danks et London, 2017), notamment au travers des jugements de valeur effectués, des erreurs potentielles et des métriques utilisées qui ne peuvent représenter toutes les dimensions d'une analyse (Suresh et Guttag, 2019). Les généralisations trop hâtives introduisent aussi ce type de biais (Suresh et Guttag, 2019). Typiquement, les biais de validation peuvent être observés lorsqu'une personne chargée du développement d'un produit manque de recul pour analyser objectivement les conséquences que peuvent avoir le produit développé. D'autres facteurs peuvent influencer l'interprétation du chercheur(euse) développant l'application, notamment des confusions entre corrélations et liens de causalité, des biais de financement pouvant par exemple conduire le bénéficiaire du financement à ne pas présenter de résultats négatifs à l'organisme responsable du financement ou encore des biais d'observateur, qui correspondent au fait que le chercheur(euse) pro-

jette ses attentes dans les résultats obtenus) (Mehrabi, Morstatter, Saxena, Lerman et al., 2021).

1.3.4 Biais dans le déploiement

Avec le déploiement de l'application, plusieurs biais trouvant leurs racines dans les étapes précédentes (spécification, données, modélisation et validation) peuvent être amplifiés. Par exemple, une mauvaise spécification du problème peut conduire à une inadaptation entre les valeurs prédites et l'interprétation de ces valeurs (prédire le succès scolaire d'un(e) étudiant(e) sans toutefois savoir comment interpréter ce qu'est le «succès») ou encore un mauvais échantillonnage des données peut conduire à des résultats erronés. Les biais dans le déploiement des modèles proviennent aussi de l'inadéquation du modèle à l'évolution temporelle des distributions, puisque les données d'entraînement ne représentent qu'un instant figé d'une réalité qui évolue. Sans mise à jour et vérification régulière, le modèle déployé ne sera pas en mesure de prédire correctement les décisions. De même, des biais peuvent provenir du fait que le modèle soit incapable de traiter des cas non vus durant la phase d'entraînement ou encore qu'un unique modèle soit utilisé dans des contextes différents. Par exemple, les méthodes d'amélioration de l'équité peuvent conduire à des modifications du comportement des personnes (si ces dernières ont l'impression d'être discriminées pour favoriser d'autres groupes), ou conduire à de nouvelles distributions en offrant des opportunités à des personnes qui n'en n'avaient pas par le passé. Ainsi, les modèles pré-entraînés ne peuvent faire face à ces nouvelles tendances s'ils ne sont pas mis à jour.

Des recherches récentes (Fazelpour et Danks, 2021 ; Mehrabi, Morstatter, Saxena, Lerman et al., 2021 ; Roselli, Matthews et Talagala, 2019 ; Silva et Kenney, 2018, 2019 ; Suresh et Guttag, 2019) identifient d'autres types de biais qui ne sont pas présentés dans cette thèse, notamment les biais de transfert de contexte, les biais d'omission de variable, les biais émergents, les biais temporels, les biais de changement de distribution, et bien d'autres.

1.3.5 Formes de discrimination

Il est important de rappeler que tous les biais ne sont pas forcément négatifs (Mehrabi, Morstatter, Saxena, Lerman et al., 2021 ; Romei et Ruggieri, 2014). Certains biais sont notamment utilisés pour résorber le risque de discrimination potentiel. Par exemple, les biais d'échantillonnage peuvent être utilisés pour réduire la proportion d'une population majoritaire afin que celle-ci ait la même

importance qu'une autre minoritaire. On peut identifier trois principales formes de discrimination, la *discrimination directe*, la *discrimination indirecte* et la *discrimination basée sur des statistiques*.

Pour illustrer ces types de discrimination, considérons le cadre des États-Unis où le clivage entre communautés est explicite et permet par exemple d'identifier l'origine ethnique d'une personne à partir de son code postal.

— La *discrimination directe* est définie comme l'utilisation explicite des attributs sensibles des personnes dans le processus de décision. Elle est parfois appelée traitement disparate (*disparate treatment*) (Romei et Ruggieri, 2014). Par exemple, la discrimination directe consisterait au refus d'un(e) candidat(e) à cause de ses croyances religieuses ou ses convictions politiques, comme interdit par exemple par la loi anti-discrimination [6].

— La *discrimination indirecte* est plus insidieuse et apparaît lorsque des attributs non-sensibles sont corrélés aux attributs sensibles et utilisés en intrants du modèle d'apprentissage. Dans notre exemple, le code postal peut être utilisé pour inférer des informations sur l'origine ethnique d'une personne. Par conséquent, rendre une décision en s'appuyant sur le code postal est fortement similaire au fait de prendre une décision en se basant sur l'origine ethnique de la personne. Un exemple bien connu de discrimination indirecte est la pratique du «*redlining*, qui consiste à refuser ou à limiter arbitrairement les services financiers dans certains quartiers, généralement parce que leurs habitants sont des personnes de couleur ou avec des faibles revenus (Dwork, Hardt, Pitassi, Reingold et al., 2012).

— La *discrimination basée sur des statistiques* concerne le fait que les décisions individuelles soient rendues sur la base des statistiques globales de groupes. En d'autres termes, les statistiques d'un groupe sont utilisées pour rendre une décision sur une personne appartenant à ce groupe, identifiable par ses caractéristiques explicites. Il s'agit typiquement pour les êtres humains de rendre par exemple une décision en s'appuyant uniquement sur des stéréotypes. Par exemple, de manière caricaturale, cela reviendrait à considérer que toutes les personnes de couleur sont des personnes défavorisées et potentiellement dangereuses, ou ayant un passé criminel, ou encore que toutes les personnes de certaines confessions religieuses sont des personnes d'intérêt pour la défense nationale.

Chacune de ces formes de discrimination peut se retrouver soit dans le contexte de la *discrimination*

6. https://www2.gouv.qc.ca/portail/quebec/ressourcesh?lang=fr&g=ressourcesh&sg=personnel&t=o&e=705068571

justifiable ou alors celui de la *discrimination non-justifiable*. Pour mieux cerner la *discrimination justifiable*, il faut revenir au fait que toute forme de discrimination n'est pas forcément négative. Par définition, la discrimination est le traitement différent de personnes en fonction de leurs différences d'attributs. La discrimination justifiable est celle qui est nécessaire pour des besoins métier [7]. Par exemple, lors d'une campagne de dépistage de cancer du sein, les personnes à tester sont choisies en appliquant majoritairement une discrimination basée sur le genre des personnes. De même, une entreprise pourrait justifier la discrimination qu'elle applique lors du recrutement en fonction des besoins de métier. Cette forme de discrimination est aussi appelée discrimination objective (Romei et Ruggieri, 2014). De manière intuitive, la discrimination justifiable concerne principalement la discrimination directe et celle statistique, puisque celle-ci assume la connaissance explicite des attributs sensibles de chacune des personnes, afin d'appliquer le traitement inégal, mais justifié.

La *discrimination non justifiable*, quant à elle, fait référence à toute forme de pratique produisant de manière non justifiée des résultats défavorables envers les personnes d'un groupe. Elle est aussi appelée discrimination illégale et s'oppose aux cadres légaux établis pour assurer le respect de la personne et sa dignité.

À ces types de discriminations communes en apprentissage automatique, on peut rajouter la *discrimination systémique ou institutionnelle*, la *discrimination individuelle* ainsi que la *discrimination structurelle* (Feagin et Eckberg, 1980 ; OHRC, 2022 ; Pincus, 1996) que nous n'approfondirons pas dans notre travail.

1.4 Mesures de discrimination et quantification de l'équité

Il existe plusieurs définitions de l'équité dans la littérature. À chacune de ces définitions sont associées une ou plusieurs métriques mathématiques. En conséquence, de nombreuses recherches en équité au sein de l'apprentissage automatique abordent la notion de l'équité à partir des différentes formulations mathématiques de cette notion, ainsi que de l'objectif à atteindre (Ferrer, Nuenen, Such, Coté et al., 2021 ; Mehrabi, Morstatter, Saxena, Lerman et al., 2021 ; Narayanan, 2018b ; Verma et Rubin, 2018). Dans cette section, nous présenterons les différentes définitions de l'équité avec un accent mis sur celles que nous avons utilisées dans nos travaux de recherches. Nous avons choisi de les présenter en regroupant les métriques d'équité en fonction de leur impact sur les

7. Charte des droits et libertés des personnes, chapitre 1, alinéa 20 [8]

groupes, les personnes humaines, ainsi que leur mode de réalisation (approche causale, mesurée sur les groupes ou sur les personnes uniquement, etc.).

Chacune de ces formes d'équité peut être appliquée avec ou sans connaissance des attributs sensibles. Dans le cas où l'attribut sensible n'est pas connu et exclu du processus d'apprentissage, on parle d'équité par la méconnaissance des attributs sensibles ou encore en anglais *fairness under unawareness*. Il existe toute une branche de recherche en équité qui s'intéresse à la résolution des problèmes de discrimination lorsque l'attribut sensible n'est pas connu (Gupta, Cotter, Fard et Wang, 2018 ; Hashimoto, Srivastava, Namkoong et Liang, 2018 ; Kallus, Mao et Zhou, 2022 ; Lahoti, Beutel, Chen, Lee et al., 2020), ou lorsque l'attribut est bruité (Wang, Guo, Narasimhan, Cotter et al., 2020). Dans le cas où les attributs sensibles sont connus et utilisés, on parle d'équité par la connaissance ou encore en anglais de *fairness through awareness*. L'attribut sensible est alors utilisé pour quantifier et mesurer les performances des approches pour la métrique qu'elles sont supposées améliorer.

Plusieurs études (Caton et Haas, 2020 ; Feuerriegel, Dolata et Schwabe, 2020 ; Mehrabi, Morstatter, Saxena, Lerman et al., 2021 ; Narayanan, 2018a ; Pessach et Shmueli, 2020 ; Suresh et Guttag, 2019 ; Verma et Rubin, 2018) ont analysé de manière plus ou moins approfondie les différences entre ces notions d'équité. En pratique, le choix de la notion à appliquer et des métriques associées est fortement dépendant du contexte d'application.

1.4.1 Équité de groupe

L'équité de groupe est la forme d'équité la plus développée dans la littérature, notamment parce qu'elle est plus facile à opérationnaliser et à satisfaire par les modèles (Bechavod, Jung et Wu, 2020 ; Carey et Wu, 2022 ; Mukherjee, Yurochkin, Banerjee et Sun, 2020). L'équité de groupe requiert en général l'égalité d'une statistique entre les groupes identifiés par les valeurs de l'attribut sensible. Autrement dit, elle peut être vue comme mesurant la similarité de traitement entre groupes définis par les valeurs de l'attribut sensible. Ainsi, on peut définir autant de métriques de l'équité de groupe qu'il existe de statistiques utilisées pour mesurer les performances d'un modèle entraîné. Par exemple, l'équité de groupe peut être quantifiée en utilisant la parité des taux de prédictions (positives ou négatives), la parité des exactitudes de prédiction, la parité des vrais ou faux positifs, l'égalité des scores de prédiction, la calibration égale dans les groupes, etc. Ainsi, le sens propre de l'équité de groupe quantifiée sera aussi dépendant de la statistique utilisée. Les études

suivantes (Caton et Haas, 2020 ; Mitchell, Potash, Barocas, D'Amour et al., 2018 ; Verma et Rubin, 2018) discutent de manière approfondie des métriques d'équité de groupe qui peuvent être obtenues à partir de la matrice de confusion ainsi que les hypothèses souvent implicites qui régissent leur choix.

Considérons la situation dans laquelle nous ne nous intéressons qu'à un seul attribut sensible binaire : l'attribut sensible S peut prendre les valeurs s_0 ou s_1. Considérons ensuite la décision originale $Y \in \{0, 1\}$, et sa prédiction $\hat{Y}$ obtenue par l'intermédiaire d'un classifieur entraîné à la prédire. Parmi l'ensemble des métriques utilisées pour mesurer l'équité de groupe, *la parité démographique* (équation 1.5), *l'effet disparate* (équation 1.7), *l'égalité des chances* (équation 1.8) ainsi que sa relaxation *l'égalité des opportunités* (équation 1.9), apparaissent comme les plus utilisées dans la littérature.

Définition 4 (Parité démographique (Pessach et Shmueli, 2020))**.** *Soit un modèle h et un ensemble de groupes identifiés par les valeurs $S = s_0$ et $S = s_1$. Le modèle h satisfait la parité démographique dans ses prédictions de la décision, $\hat{Y}$, si les taux de décisions positives dans chacun des groupes identifiés sont égaux :*

$$P(\hat{Y} \mid S = s_0) = P(\hat{Y} \mid S = s_1). \tag{1.5}$$

La parité démographique (définition 4) a plusieurs défauts. Premièrement, elle ne tient pas en compte la performance du classifieur utilisé ni son mode de réalisation. En effet, la métrique se calcule uniquement à partir des prédictions faites par le classifieur, et ne fait pas intervenir d'élément permettant d'évaluer la validité des prédictions réalisées dans les groupes. Ainsi, un classifieur qui prédirait toujours la même décision, quel que soit l'intrant, satisferait parfaitement une parité démographique. Concernant le mode de réalisation, un modèle de prédiction qui satisfait la parité peut être sujet au phénomène de paresse (*laziness* en anglais). Ainsi, la parité démographique pourrait être satisfaite par exemple en sélectionnant $p\%$ des personnes les plus qualifiées dans un groupe et en choisissant aléatoirement $p\%$ dans l'autre groupe (Carey et Wu, 2022 ; Galhotra, Brun et Meliou, 2017). Ensuite, si un classifieur a de bonnes performances dans un groupe et des mauvaises dans l'autre, ce classifieur pourrait satisfaire la parité démographique en réduisant ses performances dans le meilleur groupe pour les ramener au niveau des celles de l'autre groupe (Chang et Shokri, 2021). Enfin, la parité démographique peut être difficilement atteignable lorsque les groupes diffèrent significativement de taille. Par exemple, dans le cas extrême où un groupe ne contient qu'un

élément alors que l'autre en contient davantage, satisfaire la parité démographique demanderait soit 100% (ou 0%) de décisions positives. En conséquence, le modèle parfait de prédiction (exactitude de prédiction de 1) ne garantit pas la parité démographique si les taux de base de décisions positives de chaque groupe sont différents.

Satisfaire parfaitement la parité démographique peut s'avérer difficile, car cette situation est atteinte lorsque la différence des taux de décisions positive entre les groupes est égale à zéro. La parité démographique peut être relaxée en cherchant à ce que la différence des taux de décisions positives entre les groupes soit inférieur à un seuil ϵ (définition 5). On parlera alors de la *parité démographique-ϵ* (ϵ- demographic parity).

Définition 5 (Parité démographique-ϵ (Pessach et Shmueli, 2020))**.** *Soit un modèle h, un ensemble de groupes identifiés par les valeurs $S = s_0$ et $S = s_1$ et ϵ une très petite valeur positive. Le modèle h satisfait la parité démographique-ϵ si la différence des taux de décisions positives dans chacun des groupes est inférieur à ϵ :*

$$\mid P(\hat{Y} \mid S = s_0) - P(\hat{Y} \mid S = s_1) \mid \leq \epsilon \tag{1.6}$$

Définition 6 (Effet disparate (Feldman, Friedler, Moeller, Scheidegger et al., 2015))**.** *Un modèle h affiche un effet disparate si le ratio des proportions des décisions positives prédites par h, entre le groupe protégé et le groupe privilégié, est inférieur au seuil ϵ :*

$$min\Big\{\frac{P(\hat{Y} \mid S = s_0)}{P(\hat{Y} \mid S = s_1)}, \frac{P(\hat{Y} \mid S = s_1)}{P(\hat{Y} \mid S = s_0)}\Big\} \leq \epsilon \tag{1.7}$$

Lorsque $\epsilon = 0.8$, l'effet disparate (définition 6) est aussi connu sous le nom de la *règle des* 80% aux États-Unis. Cette règle est utilisée par la commission américaine pour l'égalité des chances en matière d'emploi (*Equal Employment Opportunity Commission* ou *EEOC*). Une des limites de cette métrique est qu'elle ne peut être calculée qu'entre deux groupes. Ainsi, pour la gestion de plusieurs groupes, certaines recherches (Kearns, Neel, Roth et Wu, 2017) cherchent à promouvoir l'équité en s'assurant que le seuil ϵ est respecté quelle que soit la paire de groupes choisie.

L'égalité des chances ($EqGap_y$, définition 7) fait intervenir la décision originale Y et consiste à s'assurer que la différence entre le taux de vrais positifs et de faux positifs soit inférieure au seuil choisi ϵ.

Définition 7 (Égalité des chances (Hardt, Price, Srebro et al., 2016)). *L'égalité des chances pour un modèle h entre le groupe protégé (s_0) et le groupe privilégié (s_1) est définie comme l'égalité des taux de vrais positifs, et des taux de faux positifs entre les deux groupes :*

$$P(\hat{Y} = 1 \mid S = s_0, Y = y) = P(\hat{Y} = 1 \mid S = s_1, Y = y) \,\forall\, y \in \{0, 1\}. \tag{1.8}$$

L'égalité des opportunités (définition 8), quant à elle, ne s'intéresse qu'à l'égalité des vrais positifs. Puisque ces métriques font intervenir les décisions originales, elles présupposent que ces décisions sont décorrélées de l'attribut sensible.

Définition 8 (Égalité des opportunités (Hardt, Price, Srebro et al., 2016)). *L'égalité des opportunités est une relaxation de l'égalité des chances dans laquelle on ne s'intéresse qu'à l'égalité des vrais positifs entre les groupes.*

$$P(\hat{Y} = 1 \mid S = s_0, Y = 1) = P(\hat{Y} = 1 \mid S = s_1, Y = 1). \tag{1.9}$$

Par exemple, considérons un ensemble de données avec les proportions suivantes : $P(Y = 1 \mid S = 0) = 0$ et $P(Y = 1 \mid S = 1) = 1$. Dans cette situation, il est facile de voir que les quantités $P(\hat{Y} = 1 \mid S = 0, Y = 1)$ et $P(\hat{Y} = 1 \mid S = 1, Y = 0)$ ne sont pas définies, la différence des taux de vrais positifs et celle des taux de faux positifs ne peuvent être calculées, il devient impossible de vérifier si l'égalité des opportunités et l'égalité des chances sont satisfaites. De même, un modèle de prédiction parfait sur les données biaisées obtiendrait une exactitude de prédiction de 1, avec des nombres de faux positifs et faux négatifs dans tous les groupes égaux à zéro, conduisant ainsi à une satisfaction de l'égalité des chances et de l'égalité des opportunités $EqGap_{y=1}$ de manière triviale. Cependant, les décisions étant biaisées, la discrimination persiste, car le modèle de prédiction reproduit parfaitement ces décisions. Cette observation peut être appliquée aux modèles performants, mais imparfaits. En conséquence, un modèle qui satisfait l'égalité des opportunités et l'égalité des chances n'est pas nécessairement exempt de biais, si les décisions originales n'en sont pas exemptes. L'égalité des chances peut aussi être relaxée en métrique α-discrimination (Γ) (équation 1.10) (Cummings, Gupta, Kimpara et Morgenstern, 2019 ; Woodworth, Gunasekar, Ohannessian et Srebro, 2017).

Définition 9 (α-discrimination (Cummings, Gupta, Kimpara et Morgenstern, 2019 ; Woodworth, Gunasekar, Ohannessian et Srebro, 2017)). *Un prédicteur rendant des décisions $\hat{Y}$*

n'est pas α-discriminant si la différence entre les taux de vrais positifs de chaque groupe, ainsi que la différence entre les taux de faux négatifs, sont bornées par la valeur α :

$$\Gamma = \underset{y \in Y}{max} |P(\hat{y} = 1 | Y = y, S = s_0) - P(\hat{y} = 1 | Y = y, S = s_1)| \leq \alpha \tag{1.10}$$

Pour l'équité exacte, la métrique doit être satisfaite pour $\alpha = 0$.

1.4.2 Équité individuelle

L'équité individuelle cherche à assurer que des personnes similaires, à l'exception des valeurs de leurs attributs sensibles, reçoivent le même traitement, c'est-à-dire des décisions identiques. Cette notion d'équité se rapporte notamment à la notion de *traitement disparate* (Barocas et Selbst, 2016 ; Zafar, Valera, Gomez Rodriguez et Gummadi, 2017) présente dans le cadre judiciaire américain. La difficulté principale pour appliquer cette notion d'équité est l'aspect subjectif de la définition de la notion de *similarité* entre les personnes.

Formellement, considérons $\mathcal{A}$ et $\mathcal{Y}$ l'espace des entrées et sorties d'un modèle d'apprentissage h (les attributs sensibles ne sont pas pris en compte).

Définition 10 (Équité individuelle (Dwork, Hardt, Pitassi, Reingold et al., 2012))**.** *L'équité individuelle (individual fairness, IF) est définie comme la continuité K-Lipschitz du modèle h en relation avec les mesures $d_{\mathcal{Y}}$ et $d_{\mathcal{A}}$, définies sur $\mathcal{A}$ et $\mathcal{Y}$:*

$$d_{\mathcal{Y}}(h(a_i), h(a_j)) \leq K d_{\mathcal{A}}(a_i, a_j). \tag{1.11}$$

Dans sa définition originale (Dwork, Hardt, Pitassi, Reingold et al., 2012), le coefficient K prend la valeur 1 et la métrique $d_{\mathcal{Y}}$ est définie de deux manières : soit par la distance statistique $d_{\mathcal{Y}} = D_{tv}$,

$$D_{tv}(P, Q) = \frac{1}{2} \sum_{a \in \Omega} |P(a) - Q(a)|, \tag{1.12}$$

soit par la métrique relative D_∞ :

$$D_\infty(P, Q) = \underset{a \in \Omega}{sup}\, log(max\{\frac{P(a)}{Q(a)}, \frac{Q(a)}{P(a)}\}) \tag{1.13}$$

où P et Q deux mesures de probabilités définies sur un ensemble fini Ω. Le choix de mesure de distance entre distributions (D_{tv} ou D_∞) influence les caractéristiques de la métrique de similarité entre les personnes. En effet, selon Dwork, Hardt, Pitassi, Reingold et al. (2012), D_{tv} est bornée entre

zéro et un, et nécessite que $d_{\mathcal{A}}$ prenne une valeur très proche de zéro pour des personnes similaires ou au contraire une valeur proche de 1 pour les personnes très différentes. D_∞ en revanche, requiert que la distance soit très petite devant un ($D_\infty << 1$), et très grande ($D_\infty >> 1$) dans le cas contraire.

La plupart des recherches s'intéressent à la définition de la mesure de similarité $d_{\mathcal{A}}$. Dwork, Hardt, Pitassi, Reingold et al. (2012) présupposent déjà l'existence et l'utilisation omniprésente d'une mesure de similarité entre les personnes, par exemple, les scores de crédit pour les demandes de financement, les combinaisons de scores de test et de résultats pour les admissions au collège, etc. Pour construire la métrique de similarité, certaines recherches définissent des mesures de distances qui sont spécifiques à chaque type d'attributs comme les attributs catégoriels, ordinaux et hiérarchisés (ensemble ordonné), etc. Ensuite, la similarité entre deux profils peut être obtenue par le calcul d'une distance finale correspondant à la moyenne de chacune des mesures de distances définies sur chaque attribut (Luong, Ruggieri et Turini, 2011), ou proposent des seuils spécifiques ϵ pour chaque distance entre attributs (John, Vijaykeerthy et Saha, 2020). Pour ce dernier cas, deux profils sont considérés similaires si chacune des distances entre attributs est inférieure au seuil ϵ défini pour chaque attribut.

En fixant le seuil à zéro pour tous les attributs non sensibles ($\epsilon_{a_i} = 0$) et à l'infini pour les attributs sensibles, la notion de similarité correspond à l'égalité des valeurs de tous les attributs, à l'exception de l'attribut sensible et de la décision. L'équité individuelle dans ce cas se rapproche du protocole de test de situation tel que défini par Bendick (2007) [9]. Ce type de similarité ressemble aux contrefactuels (Lewis, 2013)qui chercher à trouver la version du profil du personne dans un autre groupe que celui d'origine.(ex : la version correspondante au profil d'une femme dans le groupe des hommes).

Plusieurs autres approches ont été développées pour l'apprentissage de la notion de similarité. Elles s'appuient souvent sur la littérature en *metric learning* existant en apprentissage automatique. Ces approches présupposent la connaissance soit d'un sous-ensemble de données qui sont jugées simi-

9. « Dans le test de situation, des paires d'assistants(es) de recherche postulent au même poste vacant réel. Au sein de chaque paire, les caractéristiques des employés susceptibles d'être liées à la productivité d'une personne employée-telles que l'éducation, l'expérience professionnelle, les certifications professionnelles et les compétences techniques - sont rendues égales par la sélection, la formation et l'accréditation des testeurs(euses) afin qu'ils ou elles semblent également qualifiés pour les postes qu'ils recherchent. Simultanément, les caractéristiques personnelles non liées au travail sont manipulées expérimentalement en jumelant des sujets de tests qui ne diffèrent que par une seule de ces caractéristiques »

laires par un auditeur externe et libre de tout biais (Ilvento, 2019), soit d'un sous-ensemble qui capture l'injustice, dans lequel l'auditeur identifie les cas de discrimination au lieu des profils similaires (Gillen, Jung, Kearns et Roth, 2018).

Les principales approches pour quantifier l'équité individuelle utilisent soit des *mesures basées sur le voisinage*, soit des *mesures basées sur les contrefactuels.*

Les métriques basées sur le voisinage utilisent une mesure de distance unique pour toutes les personnes et quel que soit le problème à résoudre, pour identifier l'ensemble des profils proches d'un profil donné. Dans ce contexte, l'équité individuelle est quantifiée par la proportion des profils voisins du profil considéré ayant obtenu des décisions identiques à celui-ci. Les profils voisins sont déterminés en utilisant la métrique de distance, sans tenir compte de l'appartenance de groupe. Dans cette catégorie, on distingue entre autre la métrique de *consistance*, définie dans (Zemel, Wu, Swersky, Pitassi et al., 2013), se base sur une mesure de voisinage.

Définition 11 (Consistance (Zemel, Wu, Swersky, Pitassi et al., 2013))**.** *Soit $\hat{y}_n$ la décision binaire pour le profil r_n obtenue par un modèle h. La consistance de h, yNN_h, mesure la similarité des décisions rendues par ce modèle dans un voisinage de profils considérés comme similaires.*

$$yNN_h = 1 - \frac{1}{N} \sum_{r_n \in R} |\hat{y}_n - \sum_{j \in kNN(r_n)} \hat{y}_j| \tag{1.14}$$

$kNN(r_n)$ représente les l plus proches voisins du profil r_n, mesuré par une distance de similarité déterminée (par exemple la distance euclidienne).

En ce qui concerne les mesures basées sur les contrefactuels, elles nécessitent la construction ou l'identification du contrefactuel d'un profil donné pour la détermination de la similarité de traitement entre les personnes. Les contrefactuels peuvent être obtenus par l'utilisation de techniques complexes telle que FlipTest (Black, Yeom et Fredrikson, 2020) qui utilisent des mécanismes de génération de données, ou être donnés par des personnes expertes dans le domaine considéré, tel que réalisé dans le test de situation (Bendick, 2007). Ce type de mesure pourrait être mis en relation avec les mesures basées sur la causalité, ce qui présente l'avantage de pouvoir capturer efficacement la similarité entre profils, contrairement aux mesures basées sur la distance. En effet, étant donné deux profils différents, la métrique de distance n'est pas nécessairement en mesure de capturer la proximité entre des valeurs d'un attribut relativement aux valeurs d'un autre. Par exemple, bien que n'ayant pas des valeurs identiques, les deux profils présentés dans le tableau 1.2 peuvent être considérés

comme étant relativement similaires. En effet, une personne experte du domaine métier “analyste de données” pourrait les considérer similaires à cause du fait que le nombre d’années d’étude dans le domaine soit contrebalancé par le nombre d’années d’expérience.

Tableau 1.2 – Exemple de profils pouvant être jugés similaires. Par exemple, l’obtention d’une maîtrise après la licence nécessite deux années, ce qui peut s’équilibrer avec les 3 années d’expérience dans le domaine.

Âge	Niveau d’Éducation	Années d’expérience	Métier
25	Licence (Baccalauréat)	3	Analyste de données
26	Maîtrise	2	Analyste de données

1.4.3 Équité par l’amélioration des performances de sous-groupes

L’équité par l’amélioration des performances de sous-groupes s’intéresse uniquement aux performances de modèles dans tous les groupes identifiables, sans effectuer de comparaisons entre ceux-ci. Ainsi, cette forme d’équité ne s’intéresse pas à l’égalisation d’une quelconque métrique entre les personnes ou entre les groupes, mais recherche uniquement à ce que les modèles affichent de très bonnes performances pour tous les groupes (Hashimoto, Srivastava, Namkoong et Liang, 2018 ; Lahoti, Beutel, Chen, Lee et al., 2020). En conséquence, cette approche évite les écueils qui peuvent survenir par la recherche d’égalité (comme la paresse ou labaisse de performances), et admet des différences de traitement entre les groupes. Ces différences de traitement pourraient avoir des conséquences importantes sur les personnes impactées.

L’équité par l’amélioration des performances de sous-groupes repose sur la *théorie de la justice* de John Rawls (2004), qui soutient le critère ou règle *Maximin* (*maximin criterion*) pour la prise de décision. D’après cette règle, la sélection de la solution la plus appropriée parmi un ensemble de solutions possibles se fait en deux étapes. La première consiste en un classement des solutions en fonction de leurs plus mauvais résultats, c’est-à-dire que les solutions sont ordonnées en s’appuyant sur la pire des conséquences que peut avoir chaque solution. Ensuite, le choix de la solution finale se fait en sélectionnant la solution ayant le meilleur résultat parmi les pires résultats de l’ensemble des solutions. (Rawls, 2004). À cause de cela, cette forme d’équité est plus connue sous le terme de *Rawlsian Fairness* ou *équité de Rawls*.

Lahoti, Beutel, Chen, Lee et al. (2020) motivent cette forme d'équité en considérant certains contextes critiques tels que le contexte *médical*, où l'obtention de la parité d'une métrique entre groupes au détriment de l'utilité n'est pas acceptable. Les performances, au contraire, doivent être améliorées dans tous les groupes, surtout ceux dans lesquels les performances sont moins bonnes. Hashimoto, Srivastava, Namkoong et Liang (2018) positionnent cette forme d'équité dans le contexte d'une utilisation continue d'un système d'apprentissage automatique, dans lequel les personnes appartenant aux groupes ayant des performances très mauvaises auront moins d'incitatifs à continuer à utiliser le système, tandis que les personnes des groupes ayant de très bonnes performances continueraient de faire usage du système. En conséquence, un déséquilibre croissant se crée entre les quantités d'informations disponibles concernant chaque sous-population.

L'équité de Rawls présente l'avantage d'être intégré dans le contexte des autres formes d'équité. En effet, Aïvodji, Bidet, Gambs, Ngueveu et al. (2021) ont pu montrer que la prévention d'inférence de l'attribut sensible améliorent aussi les performances de prédictions dans les groupes. Puisqu'il s'agit de l'amélioration de performances, cette famille d'équité est mesurée par les performances des modèles dans les différents groupes, plus précisément leur amélioration dans les groupes identifiables. Ainsi, les métriques peuvent être le risque empirique minimal (ou résultat de la fonction de perte), l'exactitude de prédiction ou toute autre mesure d'utilité standard.

1.4.4 Équité causale

L'équité causale permet d'analyser si les attributs sensibles influencent les prédictions d'un classifieur au travers des attributs non-sensibles (Salazar, Neutatz et Abedjan, 2021), qui pourraient agir comme intermédiaires (*proxy*) pour ces attributs sensibles. L'équité causale cherche ainsi à modéliser les relations causales entre les attributs, pour découvrir des groupes ou sous-groupes subissant des différences de traitement, ainsi que les sources de biais (Caton et Haas, 2020 ; Loftus, Russell, Kusner et Silva, 2018). La connaissance de l'origine des biais permet de comprendre l'impact de leurs propagations et selon Loftus, Russell, Kusner et Silva (2018) évite les ambiguïtés posées par les méthodes s'appuyant sur les relations statistiques et les corrélations.

Les approches portant sur l'équité causale s'appuient pour la plupart sur le cadriciel proposé par Pearl (2009) : le modèle causal structurel (*Structural Causal Model)* ou *SCM*. Dans ce cadriciel, un modèle causal est défini par le quadruplet $\langle \mathbf{U, V, F, P(U))} \rangle$ dans lequel :

— $\mathbf{V}$ correspond à l'ensemble des variables observées qui constitueront les noeuds du graphe

(ex : les noeuds *GPA*, *Race*, *Sex*, *LSAT* ou moyenne des notes de la première année (*FYA*, first year average grade)).

— **U** représente les variables latentes qui modélisent l'ensemble des causes possibles des variables **V** (ex : les noeuds U_{GPA} ou U_{Race}).

— **P(U)** représente la distribution conjointe que suivent les variables **U**.

— **F** est l'ensemble des fonctions f_v qui permettent d'exprimer les valeurs de l'ensemble observé **V** en fonction des ensembles **U** et **V**.

Un exemple de graphe causal est montré dans la figure 1.1, inspiré de Kusner, Loftus, Russell et Silva (2017). De manière informelle, un lien de causalité peut être établi entre deux variables ou noeuds lorsqu'il existe au moins deux assignations de valeurs différentes sur un noeud conduisant à des distributions de probabilités différentes sur un autre (Loftus, Russell, Kusner et Silva, 2018). Les assignations de valeurs d'un ou de plusieurs noeuds du graphe sont appelé *interventions* dans le cadre du raisonnement causal. Elles peuvent forcer les variables à prendre des valeurs très peu probables afin d'étudier le comportement d'un système. Par exemple, si les interventions $U_{GPA} = 10$ et $U_{GPA} = 100$ causent des modifications sur le GPA, alors on peut établir (informellement) que U_{GPA} est l'une des causes de GPA. Ainsi, dans cet exemple, la variable U_{GPA} pourrait correspondre au niveau de connaissance global de l'étudiant.

À partir du SCM, on peut définir plusieurs effets tels que l'*effet de chemin spécifique*, l'*effet causal total* ou encore l'*effet contrefactuel* (Carey et Wu, 2022 ; Xu, Wu, Yuan, Zhang et al., 2019). Ces effets servent aussi de mesure pour l'équité causale. L'*effet de chemin spécifique (ECC)* mesure l'impact qu'à l'assignation de valeur d'un attribut sur un autre qui en est dépendant, en suivant uniquement un chemin prédéfini. Par exemple, on peut mesurer l'effet de chemin spécifique suivant le chemin $C1 : Sex \rightarrow Y$, le chemin $C2 : Sex \rightarrow LSAT \rightarrow Y$ ou encore $C3 : Sex \rightarrow FYA \rightarrow Y$. Si le chemin choisi correspond au chemin direct entre une décision et l'attribut sensible, (par exemple le chemin $C1$), on parle alors de discrimination directe.Dans le cas où le chemin choisi est un chemin indirect (chemins $C2$ et $C3$), on parlera de discrimination indirecte. L'effet se mesure sur l'attribut dépendant par la différence de probabilités obtenues avant et après l'assignation de valeurs.

L'*effet total causal*, par contre, mesure l'effet de l'assignation de valeurs suivant tous les chemins reliant les deux variables concernées. Ces chemins comprennent aussi bien les chemins directs que les chemins indirects. Cela correspond par exemple à la somme de tous les *effets de chemin spécifiques* entre Sex et Y : $C1$, $C2$ et $C3$. Dans le cas de l'effet total ou de l'effet de chemin spécifique,

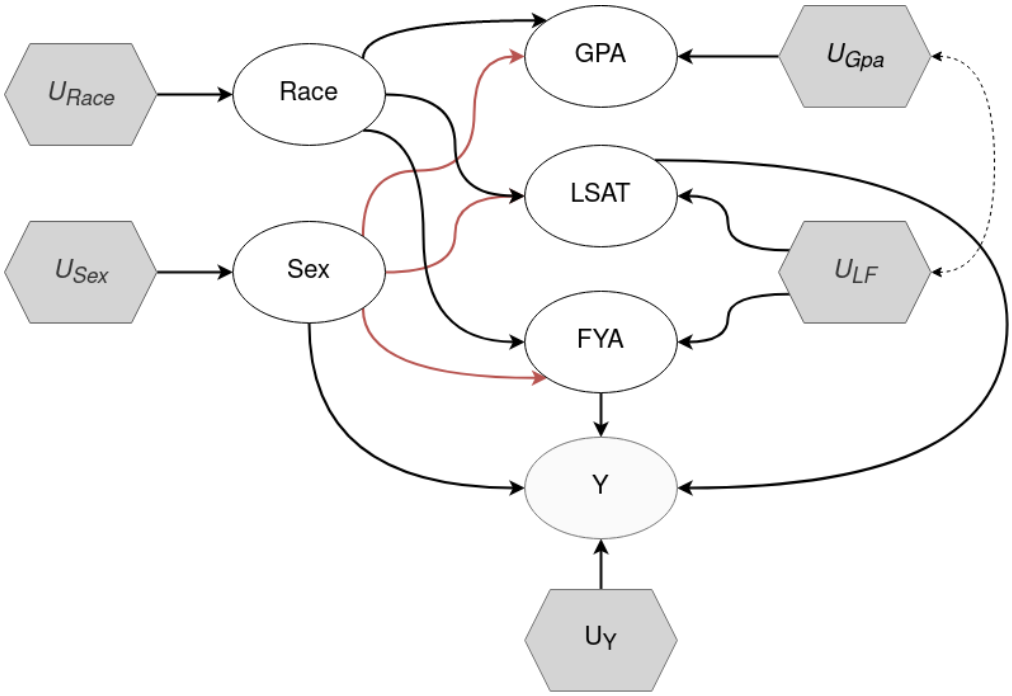


FIGURE 1.1 – Exemple de graphe causal. Les variables ayant un préfix U représentent les variables non observées influençant celles observées. Dans ce graphe, U_{GPA} est l'une des causes de GPA. Deux interventions $U_{GPA} = 10$ et $U_{GPA} = 100$ conduirait nécessairement à des distributions de valeurs différentes sur la variable GPA.

les modifications impliquées par l'assignation de valeur sur une variable sont transmises sur tous les noeuds intermédiaires entre la variable sur laquelle est réalisée l'intervention et la variable dépendante considérée. Par exemple, le changement de valeur de la variable Sex sur le chemin $C2$ impliquera aussi des changements appropriés sur le noeud $LSAT$.

Enfin, l'*effet contrefactuel*est mesuré lorsqu'un changement est appliqué sur une variable, sans être propagé sur les autres noeuds du graphe. Par exemple, considérant le profil d'une personne $Race = a$, $GPA = 3,5$, $Sex = F$ et son contrefactuel $Race = b$, $GPA = 3,5$, $Sex = F$ obtenu en intervenant sur l'attribut $Race$, l'effet contrefactuel mesure la différence de décision obtenue avec les caractéristiques originales, et celle obtenue avec le contrefactuel. Si l'effet contrefactuel est nul, alors l'équité contrefactuelle est satisfaite.

Une des limites principales de l'approche causale est la validité du modèle construit. En effet, la construction d'un graphe causal nécessite la connaissance complète du domaine, ce qui n'est pas nécessairement accessible ou réaliste comme hypothèse (CAREY et WU, 2022). La question de la validité

est encore plus importante pour les contrefactuels (Carey et Wu, 2022 ; Kusner, Loftus, Russell et Silva, 2017 ; Loftus, Russell, Kusner et Silva, 2018 ; Salazar, Neutatz et Abedjan, 2021). La connaissance partielle du domaine peut aussi rendre les effets causaux non-identifiables. Cette situation existe lorsqu'il existe deux ou plusieurs modèles causaux qui sont supportés (qui permettent d'expliquer) par les mêmes données observées, mais qui conduisent à des effets causaux différents (Wu, Zhang, Wu et Tong, 2019). Ainsi les mesures d'effet total, d'effet de chemin spécifique et les autres mesures deviendraient non explicables, c'est-à-dire non identifiable à une cause. Par conséquent, puisque les mesures d'équité causales sont basées sur ces effets, celles-ci sont aussi non identifiables (Carey et Wu, 2022).

1.4.5 Équité par la prévention d'inférence de l'attribut sensible

Puisque la discrimination est définie comme une différence de traitement entre personnes appartenant à différents groupes, l'identification de ces groupes se fait par la connaissance de l'attribut sensible S (ou des attributs sensibles). Par conséquent, sans la connaissance de cet attribut, il devient difficile d'appliquer des traitements différents entre les personnes. Cependant, l'omission de cet attribut dans l'ensemble de données n'est pas suffisante pour prévenir tout risque de discrimination à cause des corrélations entre cet attribut et les autres qui pourraient permettre d'inférer l'information sensible.

L'équité par la prévention d'inférence de l'attribut sensible cherche à empêcher l'inférence de l'attribut sensible par la suppression de l'attribut sensible ainsi que de toutes les corrélations qui pourraient permettre l'inférence de cet attribut. Puisque l'attribut sensible sera décorrélé des autres attributs, les prédictions réalisées en utilisant les données transformées vont essentiellement devenir indépendantes de cet attribut. En conséquence, certaines recherches font référence à cette approche par le terme *prédictions indépendantes des groupes*.

L'incapacité à prédire l'attribut sensible peut être mesurée à partir de *l'exactitude de prédiction* Acc d'un classifieur f entraîné à inférer cet attribut à partir de l'ensemble de données. Dans cette situation, l'exactitude de prédiction sera noté $SAcc$.

Définition 12 (Exactitude de prédiction de l'attribut sensible)**.** *Soit un classifieur f entraîné à prédire l'attribut sensible S par rapport à un ensemble de données R. Son exactitude de prédiction de l'attribut sensible se définit comme :*

$$SAcc(f, S) = \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}(s_j = f(\{y_j, a_j\})) \tag{1.15}$$

où s_j représente la valeur de l'attribut S du profil r_j.

Ainsi, $\mathbb{1}(s_j = f(\{y_j, a_j\}))$ vaut 1 si f prédit correctement s_j à partir du profil r_j. La valeur optimale (en termes de protection) de l'exactitude de prédiction qui indiquerait l'impossibilité d'inférence de S est la proportion du groupe majoritaire dans l'ensemble de données. Par exemple, pour un ensemble de données composé de 63% de la population privilégié et 37% de la population protégée, la protection optimale correspondrait à la valeur $SAcc(f, S) = 63\%$. La protection de l'inférence de l'attribut sensible peut aussi se mesurer au travers du taux d'erreur équilibré (*Balanced Error Rate* ou *BER* en anglais).

Définition 13 (Taux d'erreur équilibré)**.** *Soit un classifieur f entraîné à prédire l'attribut sensible S par rapport à un ensemble de données R. Son taux d'erreur équilibré par rapport à la prédiction de l'attribut sensible se définit comme :*

$$\begin{aligned} BER(f, S) &= \frac{1}{k}\left(\sum_{i=0}^{k-1} P(f(A, Y) \neq s_i \mid S = s_i)\right) \\ &= \frac{1}{k}\left(\sum_{i=0}^{k-1} 1 - P(f(A, Y) = s_i \mid S = s_i)\right) \\ &= 1 - \frac{1}{k}\left(\sum_{i=0}^{k-1} P(f(A, Y) = s_i \mid S = s_i)\right), \end{aligned} \tag{1.16}$$

Le BER capture la prédictibilité de chacun des groupes et prend des valeurs dans l'intervalle $[0, 1]$. À partir de l'équation 1.16, on peut observer que la valeur idéale de la protection mesurée par le BER est de $\frac{k-1}{k}$, elle correspond à la situation où les inférences réalisées ne sont pas meilleures que des choix aléatoires. En effet, un classifieur qui réalise des prédictions aléatoires peut être considéré comme un tirage aléatoire avec remise, dans lequel la probabilité de succès du tirage d'un élément du groupe s_i est donnée par sa proportion au sein de l'ensemble de donnée, $P(S = s_i)$. Puisque le tirage est aléatoire, la connaissance du groupe d'origine $S = s_i$ ne donne aucune information sur les prédictions : $P(f(A, Y) = s_i \mid S = s_i) = P(f(A, Y) = s_i)$. De même, dans le cadre d'un tirage aléatoire avec remise, la loi des grands nombres [10] nous permet d'écrire $P(f(A, Y) = s_i) = P(S = s_i)$. On obtient ainsi $\sum_{i=0}^{k-1} P(f(A, Y) = s_i \mid S = s_i) = \sum_{i=0}^{k-1} P(S = s_i) = 1$.

10. La loi des grands nombres exprime le fait que les caractéristiques d'un échantillon aléatoire d'une population se rapprochent des caractéristiques statistiques de la population considérée lorsque la taille de l'échantillon augmente

Le BER et le $SAcc$ sont complémentaires, particulièrement dans les ensembles de données ayant un déséquilibre important au niveau des tailles de groupes identifiés par les valeurs de l'attribut sensible. Dans ce type de situation, le $SAcc$ pourrait prendre une valeur élevée, ce qui pourrait laisser croire à la possibilité d'inférence de l'attribut S, tandis que le BER refléterait mieux la difficulté de prédiction de cet attribut. Cependant, le BER ne permet pas de comprendre ou d'analyser le comportement du classifieur, étant donné que les valeurs de cette métrique sont calculées en prenant la moyenne sur les groupes. Par exemple, si l'ensemble de données est composé de deux groupes ($k = 2$), dont l'un a une proportion de 85% et l'autre 15%, alors obtenir un $SAcc(f, S) = 85\%$ pourrait être obtenu de différentes manières selon que le $BER(f, S) = 0.1$ (aucune protection) ou que le $BER(f, S) = 0.5$ (le modèle prédit toujours la classe majoritaire, protection maximale), comme présenté dans la figure 1.2.

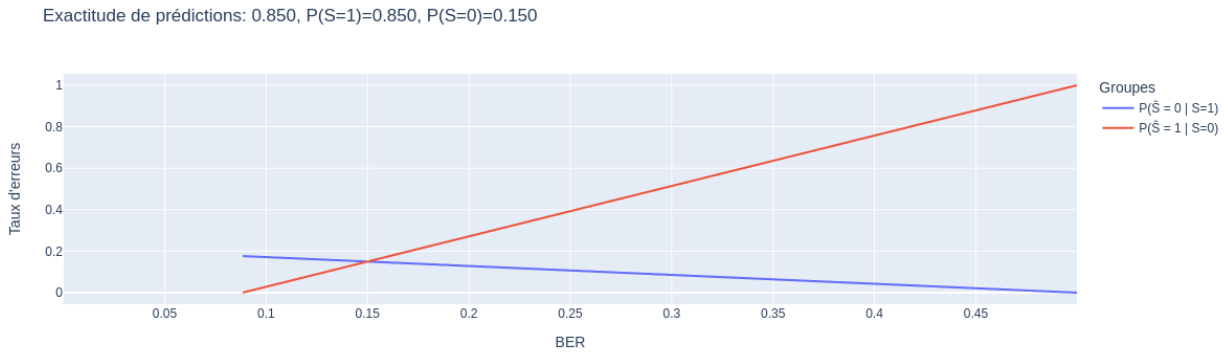


FIGURE 1.2 – Courbes représentant les valeurs possibles des proportions d'erreurs dans le contexte où le modèle de prédiction obtient une exactitude de prédiction de 0.85, et que l'ensemble de données est composé de deux groupes de proportions respectives 0.85 et 0.15. On observe les différentes valeurs des taux d'erreurs dans chaque groupe, selon la valeur du BER.

De même, un $BER(f, S) = 0.5$ (optimal) serait interprété comme l'absence de protection si le $SAcc(f, S) = 5\%$ (dans ce cas, il suffit d'inverser les prédictions du modèle pour obtenir des prédictions correctes) ou une protection efficace si $SAcc(f, S) = 85\%$. Le $SAcc$ peut ainsi être utilisé pour mieux interpréter le BER en relation avec les proportions de groupes. La protection d'inférence de

---

à l'infini. Par exemple, la loi garantit que, lorsque le nombre de tirages successifs d'une pièce sur le côté pile ou face (ou plus généralement, de tirages effectués selon une loi de probabilité) tend vers l'infini, la moyenne (ou espérance) calculée à partir des observations converge vers la moyenne réelle d'une variable aléatoire suivant cette loi.

l'attribut sensible passe ainsi par l'augmentation du BER pour le rendre près de sa valeur optimal de protection ($\frac{k-1}{k}$) ainsi que la diminution du $SAcc$ que l'on veut rendre proche de la proportion du groupe majoritaire. Enfin, à partir du BER, on peut définir l'ϵ-équité.

Définition 14 (ϵ-équité (Feldman, Friedler, Moeller, Scheidegger et al., 2015))**.** *Un ensemble de données R est considéré ϵ-équitable si pour tout mécanisme de classification f, $f : A \to S$, $BER(f(A), S) > \epsilon$.*

L'équité par la prévention d'inférence de l'attribut sensible peut être reliée à certaines notions de l'équité de groupe. En effet, à partir de l'indépendance, on peut observer une relation avec la parité démographique (*cf.* équation 1.17). La relation peut être établie parce qu'une fonction de variables indépendantes produit aussi des résultats indépendants. Ainsi, si S et A sont indépendants, alors $h(A)$ et $g(S)$ sont aussi indépendants. En posant g comme la fonction identité, alors on retrouve l'indépendance de $h(A)$ par rapport à S. Soit h une fonction définie de $A \to Y$ et $S \in \{s_0, s_1\}$:

$$S \perp A \implies \begin{cases} P(A/S = s) = P(A) \\ P(h(A)/S = s) = P(h(A)) \\ P(h(A)/S = s) = P(h(A)) \\ P(h(A)/S = s_0) - P(h(A)/S = s_1) = P(h(A)) - P(h(A)) = 0, \quad \text{parité démographique} \end{cases} \tag{1.17}$$

La relation avec l'égalité des chances peut aussi être obtenue si la décision originale est indépendante de l'attribut sensible (*cf.* équation 1.18). Cette observation suit le même principe que celui énoncé précédemment pour la parité démographique. Il suffit de considérer que la probabilité $P(h(A, Y))$ peut s'écrire comme une fonction $f(A)$, et de reprendre la même démonstration.

$$S \perp \{A, Y\} \implies \begin{cases} P(h(A)/S = s, Y = y) = P(h(A)/Y = y) \\ P(h(A)/S = s, Y = y) = P(h(A)/Y = y) \\ P(h(A)/S = s_0, Y = y) - P(h(A)/S = s_1, Y = y) = P(h(A)/Y = y) - P(h(A)Y = y) = 0 \end{cases} \tag{1.18}$$

Dans les données réelles, il arrive souvent que la décision collectée soit biaisée par rapport à l'attribut sensible. Ainsi, sans modification de cette décision, l'égalité des chances ou plus généralement toutes contraintes d'égalité entre groupes faisant appel à cette décision originale peut être difficile à satisfaire. De même, si la décision est modifiée (c'est-à-dire modifiée par rapport à sa valeur originale),

la prévention d'inférence de l'attribut sensible peut supprimer des corrélations non désirées entre S et Y de sorte que les performances du classifieur à prédire la décision dans les sous-groupes soient améliorées. Ceci peut s'expliquer par le fait que la modification de la décision peut rendre similaire les distributions conditionnelles de celle-ci entre les groupes (Aïvodji, Bidet, Gambs, Ngueveu et al., 2021). Il existe par conséquent une relation entre l'équité par la prévention d'inférence de l'attribut sensible et l'équité par l'amélioration des performances de sous-groupes.

1.5 Ensembles de données utilisés en équité

Comme mentionné dans la partie 1.3, les données sont des sources importantes de biais dans les modèles. En effet, celles-ci constituent la «matière première de l'apprentissage automatique» et leur compréhension est nécessaire à l'apprentissage de modèles permettant de résoudre des problèmes identifiés. Cependant, dans la littérature, la plupart des recherches utilisent les ensembles de données sans totalement connaître tous les aspects importants entourant la collecte et l'utilisation d'un ensemble de données.

Pour apporter plus de clarté autour de la création des ensembles de données ainsi que de responsabiliser les scientifiques dans leurs utilisations, Gebru, Morgenstern, Vecchione, Vaughan et al., 2021 proposent d'accompagner chaque ensemble par une fiche technique, similaire à ce qui se fait en médecine et en électronique, donnant des informations pertinentes suivant plusieurs axes tels que la motivation, la composition, le processus de collecte, le prétraitement ou nettoyage des données, ainsi qu'à l'utilisation prévue à la distribution et la maintenance. En se basant sur ces dimensions, une personne souhaitant utiliser les données peut évaluer leur pertinence dans différents contextes ainsi que les risques qui découlent de leur utilisation. Suivant la même logique, Holland, Hosny, Newman, Joseph et al. (2020) ont inventé une fiche descriptive adressée à un public plus expert, qui permet de comprendre les distributions et les relations entre attributs. Tout comme la fiche nutritive d'un aliment influence les habitudes alimentaires, la fiche descriptive d'un ensemble de données présente des statistiques « prêtes à l'emploi » permettant de s'assurer de l'adéquation de l'ensemble de données par rapport au problème qu'on cherche à résoudre.

Soremekun, Papadakis, Cordy et Traon, 2022 ont mené une étude sur les ensembles de données utilisés dans la littérature en équité et ont montré que les ensembles de données *Adult Census Income*, *German Credit* et *COMPAS* sont présents dans plus de 50% des articles analysés. Cette observation a été confirmée par Fabris, Messina, Silvello et Susto (2022). Ainsi, les recherches en équité

se font souvent sur des données structurées de type données tabulaires et des tâches de prédictions centrées autour des questions de finances, de vision par ordinateur et de traitement naturel de la langue (Soremekun, Papadakis, Cordy et Traon, 2022). Le Quy, Roy, Iosifidis, Zhang et al. (2022) ont conduit une analyse approfondie de quinze ensembles de données très utilisés dans la littérature en équité et ont réalisé l'analyse exploratoire de ces données. Ils (elles) ont analysé en particulier les relations entre les attributs et celles entre les attributs protégés et la décision. Les auteur(e)s montrent que les biais peuvent apparaître aussi bien au sein des données collectées que dans les prédictions des modèles et discutent notamment du fait que les problèmes d'équité sont similaires à plusieurs domaines (tel que l'éducation, la finance, etc.), mais chacun de ces domaines utilise un choix différent de l'attribut sensible ainsi qu'une définition différente de l'équité. La conséquence de ces différences est la difficulté d'évaluation des algorithmes d'équité, puisque ceux-ci reposent les mesures d'équité qui peuvent être propres à chaque domaine. Enfin, leur analyse montre aussi que les données collectées sont souvent assez anciennes (plus de 20 ans) et peuvent ne pas refléter les temps actuels. Fabris, Messina, Silvello et Susto (2022) ont réalisé une étude complémentaire à Le Quy, Roy, Iosifidis, Zhang et al. (2022) sur plus de 200 jeux de données et ont proposé la notion de fiche de données (*data brief*) qui constitue une version simplifiée de la fiche descriptive de Holland, Hosny, Newman, Joseph et al. (2020).

Ci-après, nous nous limiterons à la présentation brève de trois ensembles de données que nous avons utilisées dans le cadre de notre recherche, soit *Adult Census Income*, *German Credit* et *Lipton*, ainsi que de l'ensemble *COMPAS*.

1. Adult Census Income est un ensemble de données généré à partir du recensement de la population américaine de 1994 qui décrit la situation financière de 48842 (45222 après retrait des lignes ou des profils ne contenant pas de valeurs) personnes. Chaque profil est caractérisé par 15 attributs parmi lesquels l'attribut *genre*, ayant pour valeur «mâle» ou «femelle», est communément choisi comme attribut sensible. L'ensemble de données a été publié originellement dans le but de prédire le niveau de revenu à partir des informations du recensement. Ainsi, l'attribut de décision pour cet ensemble est «income level» qui correspond à un niveau de revenu en dessous ou au-dessus de 50000$. Adult fait partie des ensembles de données les plus utilisés en équité, notamment à cause du nombre de son nombre de profils et d'attributs.
2. L'ensemble de données German Credit est composé de profils de 1000 personnes ayant effectué une demande de crédit. Chaque profil est décrit par 21 attributs catégoriques et numé-

riques, et l'attribut sensible ici est l'*âge* alors que l'attribut de décision consiste en la qualité du client (bon ou mauvais payeur). L'ensemble a été utilisé au sein de plusieurs études telles que (Friedler, Scheidegger, Venkatasubramanian, Choudhary et al., 2019) et (Feldman, Friedler, Moeller, Scheidegger et al., 2015). Kamiran et Calders (2009) ont notamment utilisé cet ensemble de données et montré que la discrimination (mesurée par la parité démographique) est maximale lorsque l'attribut âge est binarisé avec une valeur de seuil de 25 ans.

3. Généré à partir du code de Black, Yeom et Fredrikson (2020), l'ensemble de données Lipton est un ensemble de données synthétique crée par Lipton, Chouldechova et McAuley (2018) pour étudier l'impact de la discrimination sur les approches d'amélioration de l'équité par modification des algorithmes. L'ensemble de données est composé de 2000 personnes caractérisées par les attributs *longueur des cheveux*, *experience de travail* et l'attribut sensible *genre*. La décision pour cet ensemble de données indique si une personne devrait être recrutée ou non. L'ensemble de données est généré de sorte que les attributs *longueur des cheveux* et *experience de travail* soient corrélés à l'attribut sensible *genre* tandis que la décision est basée uniquement sur l'attribut *experience de travail*. Lipton a notamment été utilisé dans la recherche (Black, Yeom et Fredrikson, 2020).

4. L'ensemble de données COMPAS est constitué de données collectées par le collectif journalistique ProPublica, données recueillies sur l'utilisation de l'outil d'évaluation des risques COMPAS dans le comté de Broward, en Floride (Angwin, Larson, Mattu et Kirchner, 2016). L'ensemble de données est composé des informations telles que le nombre de délits mineurs et le degré d'accusation de l'arrestation actuelle pour 6167 individus, ainsi que des attributs sensibles *race* et *sexe*. À chaque personne, est attribuée une décision binaire de "récidive" indiquant si elle a été réarrêtée dans les deux ans suivant la première arrestation. Nous avons suivi la procédure de prétraitement de l'ensemble de données réalisée par Friedler, Scheidegger, Venkatasubramanian, Choudhary et al. (2019).

Dans le tableau 1.3, nous présentons quelques distributions des ensembles de données. Puisque les ensembles de données Adult Census et German credit sont utilisés par une grande majorité des méthodes, il nous est facile de comparer les performances de nos approches sans avoir à nécessairement devoir ré-entraîner les méthodes de l'état de l'art. De plus, la cardinalité et la dimensionnalité de l'ensemble de données Adult en font un bon candidat pour l'utilisation des approches que nous avons développées, car celles-ci nécessitent une grande quantité de données pour correctement modéliser

TABLEAU 1.3 – Description des distributions des ensembles de données Adult Census, German Credit et Lipton. La lettre P désigne la proportion. BER Optimal and $SAcc$ Optimal correspondent aux valeurs optimales du BER et du $SAcc$ à obtenir par un classifieur pour considérer l'attribut sensible comme étant protégé.

Ensemble de données	Taille	Attributs sensibles	#groupes	P Groupe Privilégié ($S = 1$)	P Groupe Majoritaire	$P(Y = 1)$	$P(Y = 1 \mid S = s_1)$	BER Optimal	$SAcc$ Optimal
Adult Census	45222	sex	2	$0,63$	$0,63$	$0,2478$	$0,3124$	$1/2$	$0,63$
Adult2	45222	sex, race	4	$0,597$	$0,597$	$0,2478$	$0,3239$	$3/4$	$0,597$
German	1000	Age in years	2	$0,81$	$0,81$	$0,7$	$0,59$	$1/2$	$0,81$
Lipton	2000	genre	2	$0,5$	$0,5$	$0,3425$	$0,27$	$1/2$	$0,5$

les distributions des groupes. Avec German Credit, en plus de la comparaison facilitée, nous pouvons observer le comportement de nos approches dans un contexte difficile, à savoir une quantité limitée de données et un déséquilibre important entre les groupes. Concernant l'ensemble de données Lipton, son utilisation dans nos approches est motivée par le fait que celui-ci est un ensemble de données synthétique, dans lequel nous avons un contrôle complet sur les distributions. Ainsi, nous pouvons analyser et mieux observer le comportement de nos approches, en excluant certains facteurs de variations. Enfin, l'ensemble de données COMPAS n'a pas été utilisé dans nos recherches à cause du fait que les métriques de protection de l'attribut sensible sont déjà très proches de leur valeur optimale (Tableau 1.4). Toutefois, cet ensemble sera utilisé dans nos recherches futures sur la gestion de plusieurs attributs sensibles.

TABLEAU 1.4 – Distribution des attributs dans le jeu de données Compas, dans lequel le BER représente le BER le plus bas obtenus parmi l'ensemble des modèles de prédictions utilisés. De même, l'exactitude de prédiction correspond à la valeur la plus élevée.

Ensemble de données	Compass							
Attribut	Sexe		Race		Sexe-Race			
Valeurs	0	1	0	1	1-0	1-1	0-1	0-0
Proportions	0,1902	0,8097	0,6593	0,3406	0,5471	0,2626	0,1123	0,0778
BER	0,4		0,415		0,5450			
$SAcc$	0,80		0,67		0,6634			

1.6 Approches d'amélioration de l'équité

Depuis quelques années, nous assistons à un développement important des approches pour améliorer l'équité. Dans cette recherche, nous avons sélectionné les critères de classification suivants pour organiser les différentes recherches dans le domaine de l'équité : le contexte d'équité, le type d'approche, la gestion des attributs et le scénario d'évaluation.

1.6.1 Critères de classification des recherches menées en amélioration de l'équité

Contexte d'équité : Connaissance ou non de l'attribut sensible. Certaines techniques d'amélioration requièrent la connaissance l'attribut sensible alors que d'autres peuvent procéder sans sa connaissance explicite. Nous distinguerons ainsi l'équité en *situation ou contexte informée* (*fairness aware*) où les attributs sensibles sont connus (Hashimoto, Srivastava, Namkoong et Liang, 2018 ; Lahoti, Beutel, Chen, Lee et al., 2020) et l'équité en *situation ou contexte non informée* (*fairness unaware*) où ces attributs ne sont pas connus explicitement (Mohammadi, Sivaraman et Farnadi, 2022 ; Salazar, Neutatz et Abedjan, 2021 ; Xu, Wu, Yuan, Zhang et al., 2019). Ainsi, dans la situation informée, une approche d'amélioration de l'équité doit connaître le genre d'une personne pour savoir comment «orienter» la classification pour cette personne. Par contre, dans la *situation non informée*, la désignation explicite de ces groupes n'est pas nécessaire et il est possible d'améliorer l'équité en ne considérant que les attributs $\{A, Y\}$.

Enfin, entre ces deux situations, on peut aussi distinguer l'équité en *situation informée et bruitée* (*noisy awareness*) dans laquelle on considère que les valeurs de l'attribut sensible ne représentent pas la vraie distribution des données, mais une distribution intermédiaire. Cette distribution intermédiaire peut être par exemple le résultat de l'application de méthodes d'anonymisation, être due un refus des personnes de révéler leurs informations sensibles, ou encore à des fausses données générées par des personnes malveillantes qui abusent du système pour essayer d'en tirer le maximum de profit.

Type d'approche pour l'amélioration de l'équité. Le type d'approche utilisé est l'un des critères les plus communément utilisés dans les articles essayant de caractériser les approches (Friedler, Scheidegger, Venkatasubramanian, Choudhary et al., 2019 ; Mehrabi, Morstatter, Saxena, Lerman et al., 2021 ; Soremekun, Papadakis, Cordy et Traon, 2022). Les trois grandes familles d'approches existantes sont le *prétraitement des données* (Aïvodji, Bidet, Gambs, Ngueveu et

al., 2021 ; Ruoss, Balunović, Fischer et Vechev, 2020 ; Salazar, Neutatz et Abedjan, 2021 ; Xu, Wu, Yuan, Zhang et al., 2019), le *post-traitement* (Kim, Ghorbani et Zou, 2019 ; Lohia, Ramamurthy, Bhide, Saha et al., 2019) et la *modification d'algorithme* (Hashimoto, Srivastava, Namkoong et Liang, 2018 ; Kusner, Loftus, Russell et Silva, 2017 ; Lahoti, Beutel, Chen, Lee et al., 2020 ; Yurochkin et Sun, 2020).

Prétraitement des données. Ces approches modifient les données avant leur utilisation pour des tâches successives d'analyse de données. L'objectif principal est de modifier l'ensemble de données afin d'enlever toutes les caractéristiques (attributs sensibles et corrélations) qui ne respecteraient pas des contraintes spécifiques d'équité afin d'améliorer le jeu de données avant son utilisation dans le pipeline de l'apprentissage automatique. Par exemple, certaines techniques vont supprimer les corrélations existantes entre l'attribut sensible et les autres attributs afin de réduire les risques de discrimination alors que d'autres vont associer un poids à chaque profil qui sera pris en compte dans le processus de classification des données pour favoriser une métrique d'équité de groupe.

Un des bénéfices des approches de prétraitement dites "génériques" est la diversité des tâches subséquentes pour lesquelles les données modifiées peuvent être utilisées : les données modifiées ne sont pas liées à une tâche d'apprentissage particulière. Ces approches essayent souvent d'introduire le moins possible de perturbations et de préserver le même espace de représentation que les données originales. À l'inverse, les approches de prétraitement optimisées pour une tâche spécifique (*prétraitement spécifique*) ont pour objectif de réduire la discrimination en changeant l'espace de représentation tout en préservant l'utilité par rapport à cette tâche précise. Elles n'offrent donc aucune garantie sur d'autres tâches. Parmi les approches de prétraitement, on peut distinguer celles qui sont basées sur les *méthodes causales*, les *méthodes de perturbations*, l'*échantillonnage*, la *pondération* et la *transformation*.

— Les méthodes *causales* s'appuient sur la structure causale de l'ensemble de données, et suppriment les relations entre les noeuds sensibles et les autres attributs afin de respecter les contraintes d'équité. La causalité provient de la découverte des relations causales entre les attributs, par exemple par l'utilisation du graphe causal. Par exemple, l'approche (Xu, Wu, Yuan, Zhang et al., 2019) s'appuie sur la technique de *CausalGan* (qui permet de générer les données tout en respectant le graphe causal des données) et introduit un discriminateur dont la tâche est de supprimer les dépendances aux attributs sensibles dans la génération.
— Les méthodes de *perturbations* consistent en l'introduction de bruits (ou de modifications)

dans les données pour satisfaire la notion d'équité choisie. Le bruit permet de modifier les données tout en préservant l'espace de représentation des données ainsi que le domaine. Par exemple, l'approche proposée par Aïvodji, Bidet, Gambs, Ngueveu et al. (2021) consiste en l'utilisation d'un auto-encodeur couplé à un discriminateur pour introduire le minimum de perturbations dans l'ensemble de données, tout en empêchant la prédiction de l'attribut sensible à partir des données bruitées. L'auto-encodeur permet de préserver l'espace des données et de minimiser le bruit introduit. Parmi les méthodes de perturbations, on peut aussi retrouver celle de relabellisation, qui modifient la décision dans l'ensemble de données pour favoriser l'équité (Kamiran et Calders, 2012).

— Concernant l'*échantillonnage*, le prétraitement des données consiste en la sélection stratifiée des profils de l'ensemble de données de sorte que le nouvel ensemble de données préserve la diversité et que les métriques d'équité soient satisfaites. L'une des premières approches d'équité mettant en oeuvre l'échantillonnage est celle de Kamiran et Calders (2012). Leur approche utilise un premier classifieur pour obtenir les profils du groupe sensible étant proche de la frontière de décision et ayant obtenu une décision négative (groupe $G_{s=0,y=0}$), ainsi que les profils du groupe privilégié proche de la frontière et ayant obtenu une décision positive (groupe $G_{s=1,y=1}$). L'échantillonnage consiste à tirer aléatoirement la proportion $\frac{P(Y=y)}{P(Y=y/S=s)}$ dans chaque groupe $G_{s,y}$.

— La pondération. Ce type d'approche associe des poids aux profils de l'ensemble d'entraînement pour favoriser l'équité. Les poids seront utilisés durant la phase d'entraînement d'un modèle, et désigneront les données sur lesquelles le modèle doit faire le moins d'erreur de classification. Pour cette raison, cette approche se trouve à la frontière entre les méthodes de prétraitement et celle de modification d'algorithme. Par exemple, Kamiran et Calders (2012) proposent, en plus de leur méthode d'échantillonnage, une méthode de pondération basée sur la probabilité conjointe de la décision et de l'attribut sensible ($\frac{P(Y=y)}{P(Y=y/S=s)}$) pour déterminer le poids de chaque profil avant d'entraîner un classifieur.

— Les méthodes de *transformation* changent d'espace de représentation des données. Cette nouvelle représentation minimise les bris d'équité tout en favorisant l'utilité sur une ou plusieurs tâches particulières. Madras, Creager, Pitassi et Zemel (2018) proposent d'encoder les données vers une nouvelle représentation indépendante de l'attribut sensible, et permettant de maintenir un maximum d'utilité mesurée par la prédiction de la décision à partir de la nouvelle représentation.

Post-traitement des sorties. Les approches de post-traitement sont appliquées sur la sortie des algorithmes d'apprentissage, autrement dit les modèles appris. Plus précisément, une approche de post-traitement prend en entrée les prédictions réalisées par le modèle et applique une transformation sur celles-ci de sorte que les nouvelles prédictions obtenues respectent les contraintes d'équité désirées. Ces approches ont l'avantage de ne pas nécessiter la connaissance du modèle sous-jacent utilisé (autrement dit, on est dans un contexte de boîte noire), simplement les prédictions du modèle ainsi que la connaissance des attributs sensibles. Cependant, certaines recherches ont montré (Woodworth, Gunasekar, Ohannessian et Srebro, 2017) que les approches de post-traitement peuvent être sous-optimales, c'est-à-dire qu'elles réduisent considérablement les performances de prédiction des modèles. En effet, ils considèrent la situation dans laquelle Y détermine A et S, avec $P(A, S) = P(A) \times P(S)$, mais les corrélations entre Y et S sont plus importantes que celles entre Y et A. Par exemple, considérons l'exemple présenté dans la figure 1.3, dans lequel la décision (le métier de *pompier*) détermine le *salaire* et l'attribut sensible *sexe*. La décision est fortement corrélée au genre (on trouve beaucoup plus de personnes de sexe masculin dans ce métier) et aussi au salaire (puisque les sapeurs-pompiers sont des fonctionnaires). Toutefois, cette corrélation avec le niveau de salaire est moins apparente que celle avec le genre, car il peut exister plusieurs catégories de fonctionnaires avec les mêmes grilles salariales. Le *genre* et le *salaire* sont décorrélés dans notre exemple.

Un modèle de prédiction h entraîné sans regard pour la discrimination sur un ensemble de don-

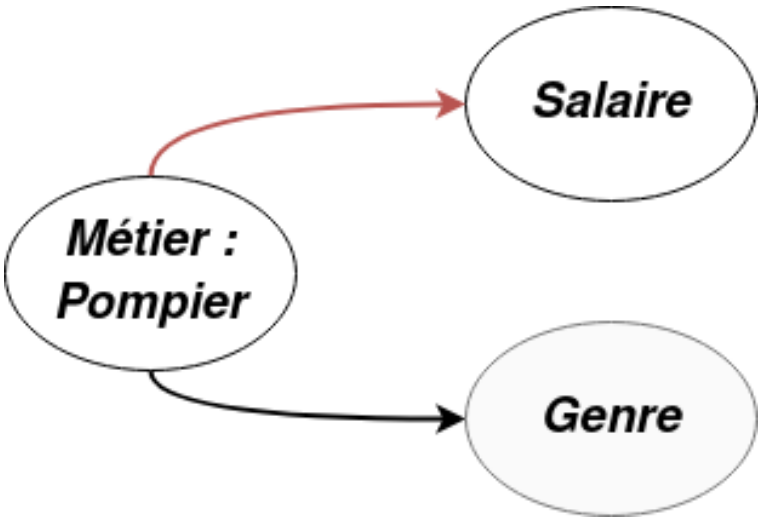


Figure 1.3 – Exemple d'illustration des limitations du post-traitement. La décision *Métier* qui prend la valeur *pompier* est corrélé au *salaire* et au *Genre*, ces derniers sont décorrélés entre eux.

nées suivant ce modèle de corrélations pourrait ne considérer que l'attribut sensible S, c'est-à-dire $h(A, S) = h(S) : h(Salaire, Genre) = h(Genre)$. Dans cette situation, réaliser un post-traitement pour améliorer l'équité reviendrait à ce que le modèle post-traité retourne une constante quelque soit S. Ce modèle est ainsi sous-optimal comparativement à un modèle entraîné pour réduire la discrimination, car le modèle évitant la discrimination n'utiliserait que les attributs A. Le post-traitement peut être réalisé par des approches basées sur la *relabellisation*, du *seuillage* ou de la *calibration*.

Modification d'algorithme. Les approches de modification d'algorithme se différencient des approches de prétraitement et de post-traitement par le fait qu'elles modifient directement l'algorithme d'apprentissage, plutôt que les données en entrée ou la sortie de l'algorithme. La modification d'algorithme va ainsi s'assurer que l'entraînement de modèle se fait en cherchant à optimiser l'équité en même temps que d'autres objectifs comme la précision de la prédiction. Les principales méthodes utilisées pour la modification d'algorithme sont l'*apprentissage antagoniste*, la *régularisation*, l'*optimisation sous contrainte*, ainsi que la *pondération*.

En plus de ces trois grandes familles d'approches, il existe des approches flexibles qui prennent la forme de *méta-algorithme* qui définissent des contraintes d'équité et font ensuite appel à d'autres méthodes de la littérature susmentionnée pour résoudre leur problème. Par exemple, dans l'approche de Lamy, Zhong, Menon et Verma (2019), les auteurs considèrent la distribution de l'attribut sensible comme étant bruitée, de sorte qu'elle correspond à une combinaison linéaire des distributions de cet attribut dans chaque groupe protégé. L'amélioration de l'équité se fait dans un premier temps par la détermination du coefficient permettant d'exprimer les bornes en équité (parité démographique et égalité des chances) sur la distribution bruitée en fonction de celle non bruitée. Ensuite, les auteurs utilisent d'autres méthodes de la littérature telle que (Agarwal, Beygelzimer, Dudík, Langford et al., 2018) qui font de l'optimisation sous contraintes pour améliorer l'équité et transforment les résultats obtenus en fonction des bornes exprimées. L'approche de Lamy, Zhong, Menon et Verma (2019) est aussi caractérisée de méta-algorithme, car elle construit un nouvel algorithme en utilisant comme briques de base de algorithmes existants dans la littérature. Un survol de ce type d'approches se retrouve dans l'étude suivante (Caton et Haas, 2020).

Gestion des attributs. La plupart des métriques et approches développées pour réduire la discrimination sont limitées à la gestion de deux groupes (c'est-à-dire à avoir un attribut sensible binaire). Par contraste, les méthodes permettant de tenir compte de plusieurs attributs sensibles se foca-

lisent sur le concept d'intersectionnalité, qui fait référence à un traitement différent des personnes en fonction de leur valeur d'attributs sensibles, dont les effets sont amplifiés sur certaines catégories de personnes se trouvant à l'intersection de plusieurs de ces attributs [11] (Foulds, Islam, Keya et Pan, 2020). Pour ce critère, nous nous référons à la description de l'environnement expérimental des approches et analysons le nombre de groupes pris en compte dans leurs démonstrations. Ainsi, certaines approches développées dans le cas binaire peuvent être étendues sur des groupes multiples, tel que mentionné par les concepteurs de l'approche. Cependant, si lors des expérimentations menées, l'approche proposée ne fonctionne qu'avec un attribut binaire, nous considérons l'approche comme binaire.

Scénario d'évaluation. Afin de clarifier le contexte d'évaluation et d'utilisation d'une approche, nous spécifions aussi le scénario d'évaluation de cette approche. Pour cela, nous nous appuyons sur les scénarios introduits dans (Aïvodji, Bidet, Gambs, Ngueveu et al., 2021), à savoir *la publication des données*, *l'entraînement d'un classifieur équitable* et *l'assainissement local* que nous décrivons ci-après.

— Publication de données. Le scénario de publication de données est typiquement celui correspondant à la plupart des techniques de prétraitement des données, où les données sont modifiées pour être ensuite réutilisées pour d'autres tâches ou analyses. Dans ce scénario, l'attribut de décision n'est pas connu d'avance, ce qui en fait un scénario similaire à celui de publication des données dans la protection des renseignements personnels, où les données sont anonymisées par une entité qui connaît l'ensemble de données originales avant d'être publiées.

— Entraînement d'un modèle équitable. Dans ce scénario, l'approche est développée dans le but d'être utilisée pour faire de la classification ou de la régression sous contraintes d'équité. Il s'agit du scénario le plus courant dans la littérature, particulièrement pour les méthodes de modification d'algorithmes et de post-traitement. Pour les approches de prétraitement des données qui sont celles étudiées principalement dans cette thèse, ce scénario consiste à l'en-

11. «*Dans sa forme générale, l'intersectionnalité met l'accent sur le fait que les systèmes d'oppression intégrés dans la société entraînent des désavantages systématiques le long de dimensions croisées, qui incluent non seulement le genre, mais aussi la race, la nationalité, l'orientation sexuelle, le statut de handicapé et la classe socio-économique. Ces systèmes sont imbriqués dans leurs effets sur les personnes à chaque intersection des dimensions concernées.*»

traînement d'un modèle sur les données modifiées afin d'obtenir un modèle non-discriminant. Ce modèle sera utilisé pour prédire les décisions en utilisant les données non modifiées. Ainsi, des modifications non désirables qui pourraient résulter de l'application du prétraitement des données seraient évitées.

— L'assainissement local (Aïvodji, Bidet, Gambs, Ngueveu et al., 2021 ; Boutet, Frindel, Gambs, Jourdan et al., 2021) est un scénario d'évaluation qui suppose que la méthode d'amélioration de l'équité sera appliquée localement par une personne, de manière privée et de sorte que les attributs sensibles ne soient connus que de lui. Dans ce scénario, l'approche transforme le profil privé de la personne utilisatrice et cette version modifiée ou transformée peut ensuite être utilisée à d'autres fins. Par exemple, l'approche peut être utilisée lors d'un sondage, où l'entité responsable du sondage met à disposition des personnes l'outil pour leur permettre répondre au questionnaire de sondage de manière privée. L'outil transforme les données des personnes de sorte que les valeurs des attributs sensibles soient masquées, avant d'envoyer la version transformée des données pour la collecte. Ainsi, seule la personne ayant rempli le questionnaire connaît les valeurs de ses attributs sensibles, car aucune autre entité n'y a eu directement accès, ce qui limite les risques de fuite d'informations. Ce scénario est similaire à celui de *la publication des données*. Il diffère toutefois en termes des hypothèses de confiance puisqu'il n'est pas nécessaire pour les personnes de confier leurs données brutes (non transformées) à l'entité responsable de la collecte des données et de leur analyse.

Ensembles de données. Le dernier critère de classification est celui qui spécifie les ensembles de données utilisées pour la validation des résultats. Ceci permet de spécifier les domaines d'applications des techniques développées, leur contexte d'utilisation ainsi que leur applicabilité à d'autres domaines.

1.6.2 État de l'art des approches d'amélioration de l'équité

Dans cette partie, nous présenterons succinctement quelques approches utilisées pour l'amélioration de l'équité. Plus précisément, la littérature dans le domaine de l'équité étant très vaste, nous nous limiterons à la présentation de quelques approches récentes, en fonction de leur contexte d'équité et de la famille d'approche concernée. Notre objectif est de présenter, pour la plupart des branches de notre arbre de critères, des exemples d'articles scientifiques récents. Nous commencerons par présenter quelques approches de l'équité dans un contexte non informé, puis nous présenterons celles

de l'équité dans un contexte informé permettant d'améliorer l'équité individuelle, l'équité causale, l'équité de groupe et l'équité par la prévention d'inférence de l'attribut sensible.

Équité sans connaissance de l'attribut sensible. En considérant l'équité de Rawls, Hashimoto, Srivastava, Namkoong et Liang (2018) ont démontré que même lorsque l'erreur empirique globale d'un modèle construit est minimale, elle peut être très élevée dans certains groupes. Ceci serait dû au fait que les fluctuations aléatoires (variation de l'exactitude de prédiction due aux caractéristiques des données) pourraient amener le modèle à favoriser certains groupes. En conséquence, le groupe favorisé verrait sa proportion augmenter car seuls les membres de ce groupe, où le taux d'erreur est bas, trouve le modèle utile. Les disparités peuvent être dès lors amplifiées, puisque le modèle n'est pas en mesure d'apprendre correctement sur les autres groupes, étant donné leur faible proportion.Pour limiter cet effet, les auteurs exploitent des techniques d'apprentissage automatique (robustesse distributionnelle) permettant d'améliorer les performances des modèles dans tous les groupes dont la taille serait supérieure à un seuil donné, au travers de la minimisation du risque empirique sur le pire cas (le pire des groupes). Leur approche présente l'avantage de s'abstraire de la définition explicite des attributs sensibles. Cependant, Lahoti, Beutel, Chen, Lee et al. (2020) considèrent que l'utilisation de la robustesse distributionnelle peut conduire à l'apprentissage de bruit. Plus précisément, ils ont enquêté sur la tâche d'amélioration des performances pour les sous-groupes identifiables par une fonction d'indication dans le contexte non informé. Les auteurs proposent une méthode d'amélioration des performances qui consiste en la modification de la fonction d'indication pour qu'elle indiquant les régions où l'erreur empirique d'entrainement est la plus large. L'apprentissage est ainsi modélisé comme un jeu à deux joueurs où le premier identifie les régions maximisant l'erreur empirique (en cherchant à identifier le plus de groupes où l'erreur est grande) tandis que le second minimise cette erreur globale. Suivant la même direction, Kim, Ghorbani et Zou (2019) proposent une approche de post-traitement pour améliorer les performances de sous-groupes. Dans un premier temps, cette approche découpe les données de l'ensemble de validation qu'ils considèrent comme des données d'audit (afin de garantir la non-réduction des performances du modèle) en fonction des probabilités de décision du modèle. Cette étape se déroule lors de l'entraînement d'un modèle considéré comme *auditeur* dont la tâche sera d'identifier les sous-populations dans lesquelles le modèle est peu performant. Enfin, ce modèle auditeur est intégré au modèle initial afin de détecter les sous-populations dans lesquelles le taux d'erreur de prédiction est supérieur à un seuil donné.

Équité avec connaissance de l'attribut sensible. Dans le contexte des approches de prétraitement, Ruoss, Balunović, Fischer et Vechev (2020) proposent d'améliorer l'équité individuelle par la construction d'une nouvelle représentation des données dans laquelle toutes les personnes similaires sont transformées vers une identique représentation z , et à partir de laquelle l'utilité dans la prédiction d'une tâche de classification est maintenue. La transformation permet ainsi de s'abstraire des corrélations et des propriétés qui rendraient difficile l'obtention et le maintien de l'équité individuelle dans l'espace de données original. Dans leur nouvelle représentation, les personnes similaires sont distantes d'au plus δ. L'équité individuelle est assurée par la non variation de la prédiction de la décision aux perturbations (d'au plus δ) de la représentation z. La notion de similarité dans cette approche est définie par un ensemble de contraintes logiques données par la personne utilisatrice. Par exemple, deux personnes de la catégorie *jeune* sont considérées similaires si l'attribut âge est inférieur à un certain seuil, et les attributs numériques de leurs profils appartiennent à la distribution des jeunes. Pour arriver à entraîner les modèles avec des contraintes logiques, cette approche se base sur le travail précédent de Fischer, Balunovic, Drachsler-Cohen, Gehr et al. (2019).Les mêmes auteurs proposent aussi un mécanisme de certification de l'équité individuelle, consistant, pour un profil donné r, à trouver la plus petite distance ϵ permettant de contenir tous les profils similaires à r et de s'assurer de l'invariance de la décision sur l'ensemble de ces profils.

Approches de modification d'algorithme pour l'équité individuelle. Yurochkin, Bower et Sun (2019) formulent l'équité individuelle comme la non-variation de la fonction de perte aux mutations des intrants, et ont conçu une approche modifiant l'entraînement des algorithmes d'apprentissage par l'ajout de la robustesse pour assurer cette équité. À partir de cette définition, ils proposent aussi une méthode d'audit d'équité individuelle par la recherche du profil maximisant la fonction de perte tout en respectant la contrainte de similarité. Pour implémenter leur approche, ils suggèrent d'apprendre des mesures de similarité à partir des données (Mukherjee, Yurochkin, Banerjee et Sun, 2020).
Yurochkin et Sun (2020) ont mis au point la notion d'équité individuelle distributionnelle et soutiennent que celle-ci peut être imposée durant l'entraînement des algorithmes par des techniques d'amélioration de la robustesse (méthodes adverseriel ou régularisation). De plus, ils soutiennent que leur approche formalise les tests de situations très utilisés dans la détection de discrimination à l'embauche aux États-Unis, dans la mesure où l'évaluation de la discrimination repose sur la différence dans les taux d'acceptations (ou de rejets) obtenues d'une part avec des profils de personnes et d'autre part avec les versions contrefactuelles de leurs profils.

Dans la même lignée, Benussi, Patane, Wicker, Laurenti et al. (2022) décrivent une approche de certification de l'équité individuelle (ϵ, δ) ainsi qu'une méthode d'entraînement des réseaux de neurones pour satisfaire cette équité. La certification passe par la recherche de l'écart maximal dans les décisions possibles pour une paire de profils distant d'au plus ϵ pris dans l'ensemble de données. Un modèle est certifié $(\epsilon, \delta) - IF$ si l'écart maximal est inférieur à δ. En ce qui concerne la méthode d'entraînement des modèles de réseaux de neurones, la contrainte d'équité individuelle est intégrée sous forme de régularisation. En conséquence, l'objectif dans l'apprentissage est de minimiser l'erreur de prédiction tout en minimisant tous les écarts maximums δ_{max_i} possibles pour chacun des profils r_i.

Approches de post-traitement pour l'équité individuelle. Les techniques de post-traitement des données pour satisfaire l'équité individuelle ne sont pas nombreuses. Une de ces méthodes, proposée par Lohia, Ramamurthy, Bhide, Saha et al. (2019), consiste à satisfaire l'équité individuelle par la modification des données de l'ensemble de test sur la base d'un détecteur de biais. L'approche construit, pour un modèle entraîné, un ensemble de données auxiliaires à partir de l'ensemble de validation. Cet ensemble auxiliaire est composé de profils du groupe protégé affichant des décisions différentes lorsque la valeur du groupe d'appartenance est modifiée. Le détecteur de biais est ensuite entraîné sur cet ensemble auxiliaire. Pour prédire la décision d'un nouveau profil, la prédiction finale est réalisée par le modèle sans aucune modification si ce profil appartient au groupe privilégié, ou appartient au groupe protégé et n'est pas considéré comme profil biaisé par le détecteur. Dans le cas où le détecteur prédit la décision comme étant biaisée, la prédiction est réalisée en changeant le groupe d'appartenance du profil à celui du groupe privilégié au moment de prédire la décision.

Mohammadi, Sivaraman et Farnadi (2022) ont conçu une approche de post-traitement et de modification d'algorithme pour améliorer l'équité individuelle. Le post-traitement consiste à modifier la sortie du modèle sur la base de la distribution de probabilités obtenue sur les possibles décisions à partir de profils similaires. En d'autres termes, pour un profil particulier, des profils qui partagent les mêmes caractéristiques à l'exception de l'attribut sensible sont générés. Ensuite, la décision attribuée au profil passé en entrée correspond à la décision majoritaire obtenue sur l'ensemble des profils similaires. L'approche de modification d'algorithme consiste à trouver, pour chaque échantillon d'entrée, le profil similaire qui maximise la différence de prédictions avec le profil initial. Par la suite, l'ensemble d'entraînement est augmenté du profil similaire identifié auquel on associe la décision du

profil original. Ces approches sont conçues de sorte qu'elles puissent bénéficier de l'optimisation MILP pour faciliter la recherche de profils similaires.

Approches de prétraitement pour l'équité causale. Concernant l'équité causale, Xu, Wu, Yuan, Zhang et al. (2019) ont conçu *Causal Fairness Generative Adversarial Networks* (*CFGAN*), technique de prétraitement des données pour générer des données synthétiques satisfaisant certains critères d'équité. Cette génération de données synthétiques se fait par l'intermédiaire de modèles permettant la génération des données (RAG, cf. partie 1.7), plus spécifiquement de l'approche CausalGan (Kocaoglu, Snyder, Dimakis et Vishwanath, 2017). L'approche CFGAN modifie CausalGan en y intégrant un deuxième générateur ainsi qu'un second discriminateur, obtenant ainsi un système composé de deux RAG dont les générateurs partagent les mêmes paramètres. Le premier réseau s'assure du respect de la distribution originale des données, tandis que le deuxième RAG s'assure de la génération sous contrainte des notions d'équité en équité causale.

D'un point de vue beaucoup plus théorique, Kilbertus, Rojas Carulla, Parascandolo, Hardt et al. (2017) étudient l'équité sous la loupe du raisonnement causal, et établissent à partir du graphe causal des relations entre les métriques utilisées dans le raisonnement causal certaines notions définies dans l'équité de groupe, telles que la parité démographique et l'égalité des chances. Ils proposent aussi un algorithme pour résoudre la discrimination non justifiable et la discrimination indirecte. Cette méthode se base sur des interventions à effectuer sur les variables corrélées et nécessite que celles-ci puissent être décomposées en une combinaison linéaire de facteurs liés à l'attribut sensible, ainsi que d'autres facteurs non observés.

Dans la même lignée des familles de prétraitement en vue de l'amélioration de l'équité causale, Salazar, Neutatz et Abedjan (2021) développent FairExp, qui permet d'augmenter les attributs disponibles afin d'améliorer l'utilité tout en réduisant la discrimination. Le lien avec la causalité dans cette approche provient de la description des attributs (ainsi que de leurs relations), où la personne utilisatrice catégorise les attributs comme sensibles, admissibles et inadmissibles. Les attributs admissibles sont des attributs sur lesquels l'influence de l'attribut sensible est acceptée, tandis que pour les attributs inadmissibles, le processus de construction des nouveaux attributs est utilisé pour extraire la partie utile afin d'améliorer l'utilité et réduire l'influence des attributs sensibles. Le processus d'augmentation des attributs consiste en la construction d'un nouvel ensemble de descripteurs par des opérations simples sur les attributs (telles que l'addition, l'écart type, etc.), et par l'utilisation d'un solveur pour

l'algèbre linéaire afin d'éviter les redondances dans le processus de génération. La construction des attributs est limitée par une longueur descriptive fixée pour ne pas générer des attributs de manière infinie. La longueur descriptive correspond au nombre d'opérations ou de transformations appliquées sur un attribut pour obtenir le nouvel attribut dénué de l'influence des attributs sensibles. Ensuite, tous les attributs sont filtrés en deux étapes : la première extrait les caractéristiques qui améliorent l'utilité, et la seconde filtre les caractéristiques qui augmentent l'équité.

Kusner, Loftus, Russell et Silva (2017) définissent une approche propre aux contrefactuels, qui repose sur un lemme stipulant que tout modèle réalisant des prédictions à partir d'attributs n'ayant pas d'attributs sensibles pour parents satisfait le principe d'équité contrefactuelle (cf. partie 1.4.4). À partir de cette définition, ils ont élaboré un algorithme d'entraînement qui repose sur le fait que le graphe causal ainsi que les facteurs non observés (et indépendants de l'attribut sensible) de chaque attribut soient connus. Le modèle de prédiction est entraîné sur un ensemble de données augmenté par une procédure qui, pour chaque profil, génère plusieurs différents facteurs non observés ou latents.L'entraînement se fait sans utilisation des attributs sensibles et de leurs descendants satisfaisant ainsi l'equité contrefactuelle peut-être satisfaite.

Approches d'amélioration de l'équité de sous-groupe. L'équité par l'amélioration des performances de sous-groupe est généralement utilisée en combinaison des techniques de modifications d'algorithmes. Dans ce contexte, Williamson et Menon (2019) ont élaboré une méthode minimisant la fonction de perte du modèle (considérée dans leur cas comme fonction de risque) tout en assurant que les erreurs calculées dans chaque groupe identifiable par le (ou les) attribut(s) sensible(s) soit équivalentes. Pour ce faire, ils utilisent une mesure de déviation de la fonction de perte entre les groupes, et le résultat de la mesure est utilisé pour régulariser la fonction de perte à minimiser. Les mesures de déviations utilisées dans leur article ont des propriétés étroitement liées à celles des mesures de risques connues et utilisées en finance, permettant ainsi l'adaptation de certaines de ces mesures au contexte de l'équité. Enfin, un coefficient α variant entre 0 et 1 est introduit pour contrôler le niveau de régularisation souhaitée, avec $\alpha \to 1$ impliquant le principe *maximin* (minimisation du risque dans tous les groupes, cf. section 1.4.3) tandis que $\alpha \to 0$ conduit à une optimisation sans contraintes.

De même, Balashankar, Lees, Welty et Subramanian (2019) ont proposé de modifier l'algorithme d'apprentissage afin que celui-ci atteigne le point efficient de Pareto dans chaque groupe

respectif. Le point efficient de Pareto est l'un des points du front de Pareto, qui correspond dans le cadre de leur approche à l'ensemble des solutions pour lesquelles il n'est pas possible d'améliorer les performances dans un groupe sans que cela ne diminue les performances au sein d'au moins un des autres groupe. Cette approche commence par chercher la performance optimale de chaque groupe pris séparément en entraînant un modèle différent par groupe. Ensuite, un modèle général est entraîné à minimiser la fonction de perte régularisée par la variance d'une mesure d'écart entre le modèle général et les modèles optimaux précédemment trouvés pour chaque groupe.

Approches d'amélioration de l'équité de groupe. Concernant l'équité de groupe, Agarwal, Beygelzimer, Dudík, Langford et al. (2018) ont conçu un algorithme améliorant la parité démographique et l'égalité des chances. Après avoir démontré que ces deux métriques peuvent être vues comme des cas particuliers d'un ensemble de contraintes linéaires, les auteurs proposent d'introduire ces contraintes dans l'algorithme. L'apprentissage est formulé comme un jeu entre deux joueurs dont le premier, appelé *apprenant*, cherche à minimiser l'erreur empirique d'entraînement, tandis que le second, dénommé auditeur, veut maximiser la satisfaction de contraintes d'équité. Zhang, Chu, Asudeh et Navathe (2021) proposent de modéliser les contraintes d'équité sous forme déclarative, c'est-à-dire sous la forme d'une somme sur tous les groupes de l'ensemble de données, de fonctions de même forme pour toutes les métriques et tous les groupes, mais dont les coefficients varient en fonction de la métrique à calculer et du groupe dans lequel on se place. À partir de cette forme, ils transforment l'optimisation des modèles de prédiction sous contrainte des métriques d'équité en optimisation pondérée de l'exactitude de prédiction.

Amélioration de l'équité par protection d'inférence de l'attribut sensible. L'approche proposée par Feldman, Friedler, Moeller, Scheidegger et al. (2015) est l'une des premières de la littérature à proposer la protection contre l'inférence l'attribut sensible pour l'amélioration de l'équité. Leur approche nommé (*Disparate Impact Remover, DIRM*), développée dans le contexte d'un unique attribut sensible binaire, consiste dans un premier temps en la construction des fonctions de répartition des attributs conditionnellement à chaque valeur de l'attribut sensible (les fonctions de répartitions sont contruites dans chaque groupe). Ainsi, pour chaque attribut, l'approche construit tout d'abord deux fonctions de répartitions, puis par la suite la médiane des deux fonctions. La protection de l'attribut sensible est réalisée par la translation des distributions conditionnelles de chaque attribut vers leur médiane : dans chaque groupe et pour chaque attribut, les nouvelles valeurs sont obtenues

en trouvant celles qui permettent d'obtenir sur la distribution médiane les mêmes probabilités que celles obtenues par les vraies valeurs sur la fonction de répartition originale du groupe considéré. Un coefficient λ est introduit pour contrôler le degré de translation, régissant ainsi le compromis entre l'utilité et la protection.

Zemel, Wu, Swersky, Pitassi et al. (2013) ont proposé l'approche *Apprentissage de représentation équitable ou Learning Fair Représentation (LFR)*, permettant d'apprendre une représentation équitable des données basée sur un ensemble de prototypes. La nouvelle représentation qui préserve l'exactitude de prédiction d'une décision et permet une reconstruction fidèle des profils originaux. Chaque profil original est décomposé en une somme pondérée des prototypes. La pondération de chaque prototype pour un profil donné est obtenue par une mesure de distance entre le profil et le prototype en question. La prévention d'inférence est assurée par des contraintes qui imposent le fait que chaque prototype ait la même probabilité d'identifier les différents groupes. Pour cette approche, la définition de l'ensemble des prototypes (c'est-à-dire le nombre de prototypes et leurs caractéristiques) est extrêmement importante. En particulier, le nombre de prototypes influence la qualité de la reconstruction : plus le nombre de prototypes est élevé, meilleure est la qualité de la reconstruction des données. Les caractéristiques des prototypes sont également importantes dans la mesure où les prototypes appartiennent au même espace que les données originales et pourraient typiquement agir en tant que représentants de la population. Leur choix doit donc être équilibré, étant donné que les profils sont transformés en des sommes pondérés des prototypes. Par exemple, le fait d'avoir un ensemble de prototypes proche du groupe privilégié induira une moins bonne qualité de reconstruction des données, en particulier pour le groupe privilégié, afin de compenser sa proximité avec l'ensemble de prototypes.

Edwards et Storkey (2015), dans leur article *Apprentissage de représentation équitable par des adversaires, ou Adversarial Learned Fair Representations (ALFR)* ont entraîné un encodeur à produire une représentation à partir de laquelle un modèle utilisé comme adversaire est incapable de prédire correctement l'appartenance à un groupe, mais préservant suffisamment d'information pour permettre à un décodeur et à un modèle de prédiction de respectivement reconstruire les données et prédire correctement la tâche choisie. Madras, Creager, Pitassi et Zemel (2018) ont étendu ce cadre pour satisfaire la contrainte d'égalité des chances (Hardt, Price, Srebro et al., 2016) et ont aussi exploré les garanties théoriques d'équité fournies par leur nouvelle représentation apprise. Enfin, ils évaluent aussi la capacité de la représentation à être utilisée pour différentes tâches de clas-

sification. Zhang, Lemoine et Mitchell (2018) ont conçu un prédicteur de décision permettant de satisfaire l'équité de groupe en garantissant qu'un adversaire soit incapable de déduire l'attribut sensible du résultat prédit.

Toujours dans la lignée des approches améliorant l'équité par la prévention de l'attribut sensible, on peut citer FairGan (Xu, Yuan, Zhang et Wu, 2018) et son extension FairGan+ (Xu, Yuan, Zhang et Wu, 2019). FairGan utilise un modèle pour générer des données à partir desquelles un prédicteur est entraîné à inférer l'attribut sensible. Le but est de pouvoir générer des données qui seront très proches des données originales, mais à partir desquelles il serait difficile de retrouver l'attribut sensible. À partir des données ainsi générées, un classifieur équitable peut-être entrainé et par la suite utilisé sur les données originales. FairGan+ étend FairGan en ajoutant un modèle de prédiction d'une décision préalablement choisie, permettant ainsi de maintenir encore plus l'utilité des données pour la décision à prendre. L'utilisation du modèle de prédiction permet aussi d'introduire de manière explicite des contraintes liées aux métriques d'égalité des opportunités et d'égalité des chances.

1.7 Réseaux adversériaux génératifs et équité

Plusieurs approches d'amélioration de l'équité, notamment celles qui nécessitent un équilibre entre deux objectifs divergents, s'appuient sur les Réseaux Adversariaux Génératifs ou RAG (Goodfellow, Pouget-Abadie, Mirza, Xu et al., 2014) (*Generative Adversarial Networks* ou GANs en anglais) pour trouver ces différents équilibres (Aïvodji, Bidet, Gambs, Ngueveu et al., 2021 ; Boutet, Frindel, Gambs, Jourdan et al., 2021 ; Wadsworth, Vera et Piech, 2018 ; Xu, Wu, Yuan, Zhang et al., 2019 ; Xu, Yuan, Zhang et Wu, 2018, 2019). Toutes les recherches que nous avons menées au long de notre thèse s'appuient sur ces méthodes, et exploitent leur caractéristiques, notamment la capacité de modélisation des données et la capacité de transformation d'une données en une autre avec un coût minimal. Nous introduirons par la suite le concept général des RAG afin de présenter avec plus de détails ces approches.

Les RAG sont des modèles utilisés pour modéliser une distribution inconnue, mais dont on connaît un ensemble d'échantillons représentés par l'ensemble de données. La version originale des RAG (Goodfellow, Pouget-Abadie, Mirza, Xu et al., 2014) consiste en deux modèles (générateur et discriminateur) qui optimise chacun des objectifs antagonistes :

1. Le premier modèle, appelé Générateur (G_{std}), produit des extrants Z à partir de bruits aléatoires X utilisés en intrant. L'objectif du générateur est de produire des données Z qui ne seraient pas différentiables des échantillons de la vraie distribution R, et ainsi d'approximer cette vraie distribution de forme inconnue.
2. Le deuxième modèle, appelé Discriminateur (D_{std}), mesure la distance entre les données générées Z et les vraies données R, et indique de quelle distribution (Z ou R) chacun des intrants provient.

À partir du retour du discriminateur, le générateur est mis à jour et est capable de modifier les données générées Z de sorte que la distance entre Z et R mesurée par D_{std} est réduite. Un jeu antagoniste se met alors en place entre le générateur (qui essaie de déjouer la détection entre les vraies données et celles qu'il génère) et le discriminateur qui essaie de séparer les données du générateur des vraies données. À terme, le générateur est en mesure d'approximer la vraie distribution R et permet, entre autres, son échantillonnage. Les générateurs dans les RAG permettent ainsi de modéliser des distributions inconnues ou difficiles à partir d'échantillons connus. La nouveauté des RAG par rapport aux autres formes de modèles génératifs existantes est la possibilité de mesurer la distance par rapport à la vraie distribution sans connaissance de sa forme ou définition explicite. Ainsi, grâce au discriminateur, l'approche peut s'abstraire de toute la famille des mesures de distribution nécessitant cette forme complète (divergences).

Dès leur conception, les RAG ont connu un succès important. Aujourd'hui, ils se retrouvent pratiquement dans tous les domaines faisant appel à l'apprentissage automatique, grâce à leur capacité de modélisation de distributions complexes et les avancées dans leur entraînement. Les RAG ont connus plusieurs avancées récentes telles que les modèles génératifs conditionnels (Mirza et Osindero, 2014) ou encore les WGAN (Arjovsky, Chintala et Bottou, 2017).

Les RAG conditionnels correspondent aux RAG dont l'objectif est de modéliser les distributions d'un ensemble de données conditionnellement aux valeurs d'un attribut spécifique, ceci afin de controler finement la génération de nouvelles données. Par exemple, dans un ensemble de données d'image de personnes, si l'attribut spécifique est le *la couleur de la peau*, l'objectif sera de modéliser les distributions d'images en fonction des couleurs, de sorte qu'une personne utilisatrice puisse générer des images de personnes d'une couleur qu'elle aurait choisie.

Les WGAN sont une forme de RAG utilisant le transport optimal de distributions. L'objectif du

transport optimal est de trouver le *coût* minimal nécessaire à la transformation d'une distribution P en une autre Q. La notion de coût correspond à une mesure de distance pondérée entre chacun des événements de la distribution P et ceux de la distribution Q. Par exemple, on définit la distance $p - Wasserstein$ par l'équation 1.19 :

$$W_p(P,Q) = \left(\inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}\, d(x,y)^p \right)^{1/p} \tag{1.19}$$

où $d(x,y)^p$ correspond à une mesure de distance, $\Pi(P,Q)$ l'ensemble des distributions conjointes entre P et Q et ayant P et Q pour marginales. L'équation 1.19 revient ainsi à trouver l'*infimum* de toutes les distances mesurées à partir de toutes les distributions conjointes entre P et Q. Les WGAN utilisent la distance 1-Wasserstein (équation 1.20), et essaient de la mesurer.

$$W_1(P,Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma} \left[|| \, x - y \, || \right] \tag{1.20}$$

Cependant, celle-ci est calculatoirement difficile, notamment à cause du l'ensemble des distributions conjointes qui peut-être très grand. Toutefois, la dualité *Kantorovich-Rubinstein* permet d'écrire l'équation 1.21 :

$$W_1(P,Q) = \sup_{||f|| \leq 1} \mathbb{E}_{x\sim P} \left[f(x) \right] - \mathbb{E}_{x\sim Q} \left[f(x) \right] \tag{1.21}$$

Ainsi, la mesure de $W_1(P,Q)$ peut se faire par la recherche du supremum de l'ensemble de fonctions lipschitziennes f. Cette recherche constituent l'un des objectifs du WGAN.

Pour trouver le supremum, les fonctions f peuvent-être modélisées des réseaux de neurones (appelés critiques ici), et l'objectif est de trouver les paramètres du réseau qui maximiserait l'équation 1.21. L'optimisation devient ainsi possible par différentiation et descente de gradients. La condition de Lipschitz peut-être assurée de différentes façon : par imposition d'une borne supérieure aux valeurs des paramètres du réseau de sorte qu'ils restent inférieur à un seuil choisit c (Arjovsky, Chintala et Bottou, 2017), en s'assurant que la norme des gradients reste inférieure ou égale à 1 (Gulrajani, Ahmed, Arjovsky, Dumoulin et al., 2017), ou en évitant la contrainte de Lipschitz en mesurant la divergence (Wu, Huang, Thoma, Acharya et al., 2018) au lieu de la distance.

En considérant la distribution Q comme étant une distribution modélisée par un réseau de neurones ou générateur G_θ, qui, pour chaque valeur z associerait un événement $G_\theta(z)$, minimiser l'équa-

tion 1.22 revient à la modélisation de la distribution P par le générateur G_θ :

$$W_1(P, G_\theta) = \sup_{||f|| \leq 1} \mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{x \sim G_\theta}[f(x)] \tag{1.22}$$

.

Le WGAN consiste, en conséquence, à la modélisation de la vraie distribution P par le générateur G_θ par la minimisation de 1.22, tandis que le discriminateur ou critique f cherche à mesurer le plus petit coût de transport (ou la plus petite divergence) de $G_\theta(z)$ vers P en maximisant l'équation 1.22.

1.8 Équité et protection de la vie privée

Plusieurs relations peuvent-être établies entre les concepts d'équité et de protection de la vie privée, notamment suivant quatre axes : le principe régisseur de chaque paradigme, l'apprentissage respectueux de la vie privée et l'équité, la similarité des approches et la non-réidentification.

1.8.1 Similarité entre principes régisseurs.

On désigne par renseignement personnel toute information qui concerne une personne physique et permet directement ou indirectement de l'identifier (Kashid, Kulkarni, Patankar et al., 2017). Un des objectifs de la protection de la vie privée est de donner à une personne le contrôle sur ses renseignements personnels. Ainsi, cette protection concerne non seulement la production, la collecte, l'utilisation et la destruction de ces renseignements, mais aussi les résultats obtenus par l'utilisation de ces informations. Par exemple, toutes inférences ou déductions faites sur des personnes à partir de leurs renseignements personnels sont aussi couvertes et régies par les mêmes principes [12] (Ministre de l'innovation des sciences et de l'industrie, 2020).

L'équité en apprentissage automatique se rapproche de la protection de la vie privée suivant deux orientations. Dans un premier temps, l'apprentissage automatique et la prise de décision font appel aux données générées par les personnes, ce qui amène directement aux mêmes problématiques que celles considérées en protection de la vie privée. De même, l'équité dans l'apprentissage automatique nécessite la collecte et l'utilisation d'attributs jugés sensibles qui amènent des risques de discrimination. Ces attributs sensibles peuvent être utilisés de manière explicite ou de manière indirecte

12. Défini au sein du cadre de la Loi 64.

par les corrélations existantes avec les autres attributs. Ainsi, puisque la protection de la vie privée se préoccupe aussi des inférences qui peuvent être faites sur les données, elle s'applique aussi à la collecte et l'utilisation de ces informations sensibles, indépendamment de leur forme (explicite ou implicite).

On peut ainsi établir un lien formel entre la protection de la vie privée et l'équité dans l'apprentissage automatique par la prévention d'inférence de l'attribut sensible (*cf.* partie 1.4). En effet, dans ces deux domaines, l'objectif est d'empêcher la connaissance ou l'inférence d'informations jugées importantes pour éviter leur utilisation. Ce lien est encore plus fort dans le cadre du scénario d'assainissement local qui consiste en l'application locale (sur téléphone intelligent par exemple) d'un outil de protection des informations sensibles de la personne utilisatrice pour supprimer les corrélations entre les attributs sensibles et les autres attributs avant leur utilisation pour d'autres finalités. Dans ce contexte, le choix de publication ou de maintien privé d'une information sensible est laissé à la personne concernée.

1.8.2 L'apprentissage respectueux de la vie privée et l'équité en apprentissage automatique

Le deuxième point de rencontre se situe entre l'apprentissage respectueux de la vie privée (*privacy-preserving machine learning*) et l'équité en apprentissage automatique. En effet, l'apprentissage respectueux de la vie privée cherche à s'assurer qu'un modèle entraîné est résistant à des attaques contre la vie privée telle que l'inférence d'appartenance, consiste pour un attaquant à déterminer si un profil connu a été utilisé pour entraîner le modèle qu'il attaque, ou encore les attaques d'extraction de modèles, qui consiste en l'extraction de l'architecture et des paramètres du modèle attaqué afin de construire un modèle imitant celui attaqué (Abadi, Chu, Goodfellow, McMahan et al., 2016 ; Liu, Ding, Shaham, Rahayu et al., 2021 ; Rigaki et Garcia, 2020). Ces données privées peuvent aussi constituer des informations sensibles qui ne devraient pas être utilisées lors de l'apprentissage en vue d'améliorer l'équité.

Pour répondre à ces objectifs, l'apprentissage respectueux de la vie privée développe des approches qui peuvent être catégorisées en utilisant une terminologie très proche de celle utilisée dans la problématique de l'équité en apprentissage automatique, à savoir le prétraitement des données, la modification d'algorithmes et le post-traitement. Par exemple, nous considérons la modification de l'optimisation dans la descente de gradient comme faisant partie de l'approche de modification d'algorithmes (Liu, Ding, Shaham, Rahayu et al., 2021 ; Rigaki et Garcia, 2020).

Aussi, le modèle de données dans la protection de la vie privée considère trois types d'informations : les informations identifiantes ID (qui peuvent correspondre à des identifiants directs ou encore à des quasi-identifiants, c'est-à-dire un ensemble d'attributs qui semblent anodins pris séparément, mais qui permettraient de réidentifier une personne lorsqu'ils sont priss ensemble), les attributs sensibles S tout comme ceux utilisés dans le paradigme de l'équité et les attributs non sensibles A. Le modèle de l'équité, quant à lui, considère que l'identifiant peut faire partie des informations non sensibles selon la métrique d'équité utilisée et considère aussi que les attributs non sensibles A incluent un attribut de décision Y.

1.8.3 Prétraitement des données

Les similarités entre protection de la vie privée et l'équité sont particulièrement visibles dans les approches de prétraitement des données. En effet, pour ces deux paradigmes, ces approches produisent des données transformées à partir desquelles il est possible d'imaginer une variété d'utilisations subséquentes. Par exemple, l'approche GANSan (Aïvodji, Bidet, Gambs, Ngueveu et al., 2021) (chapitre 2) que nous avons développée cherche à minimiser le nombre de perturbations à apporter à un profil de sorte que l'attribut sensible ne puisse plus être inféré à partir du profil modifié. Par la suite, GANSan a été adapté au contexte de la protection des informations sensibles issues de capteurs dans une approche appelée DYSAN (chapitre 3).

Du côté de la protection de la vie privée, Ping, Stoyanovich et Howe (2017) ont conçu une méthode de génération de données synthétiques qui assure la confidentialité différentielle appelée *DataSynthesizer*. Cette approche cherche à apprendre la distribution des données ainsi qu'un graphe bayésien modélisant les relations entre attributs afin de préserver la structure des corrélations de l'ensemble de données. Un autre module permet ensuite de générer de manière privée des données synthétiques à partir du graphe appris. Cette approche est très similaire des approches de génération de données synthétiques dans le contexte de l'équité, à savoir FairGan (Xu, Yuan, Zhang et Wu, 2018) et FairGan+ (Xu, Yuan, Zhang et Wu, 2019).

1.8.4 Similarité entre techniques d'anonymisation et méthodes d'amélioration de l'équité

Un sous-domaine de recherche important en protection de la vie privée est la non réidentification des personnes en utilisant des méthodes d'anonymisation des données. En particulier, les méthodes d'anonymisation peuvent être vues comme s'occupant de la protection des différentes lignes d'une

base de données, où chaque ligne représente une personne, tandis que les méthodes d'équité travaillent plutôt sur les colonnes de celle-ci (*cf.* figure 1.4). Plus précisément, l'anonymisation s'intéresse à empêcher que l'attribut *ID* puisse être prédit à partir des attributs *S, A et Y*, tandis que l'équité s'intéresse aux relations entre *S* et *A* et entre *S* et *Y*, comme détaillé dans la figure 1.5. À partir de ces deux figures, on peut constater notamment que les méthodes d'amélioration de l'équité et d'anonymisation peuvent avoir de l'influence les unes sur les autres.

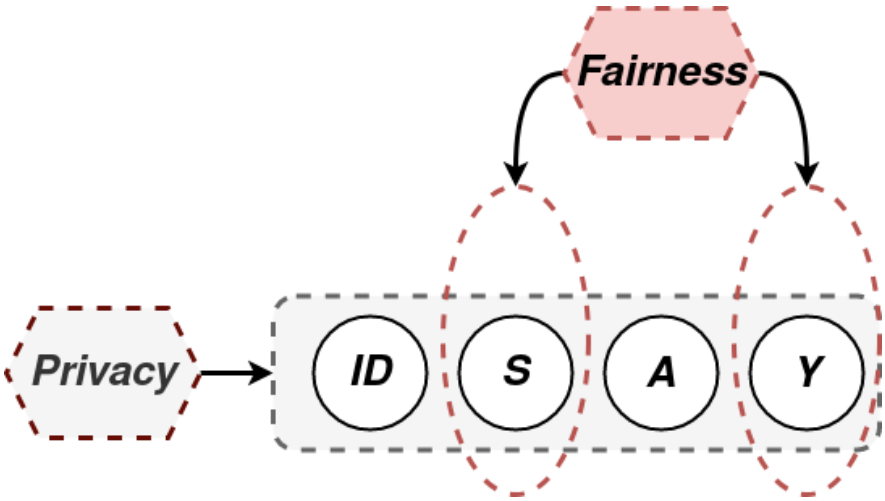


FIGURE 1.4 – Relation entre la protection de la vie privée et l'équité. La protection de la vie privée cherche à prévenir la fuite d'information tandis que l'équité cherche à limiter les risques de discrimination (MORITZ, 2013).

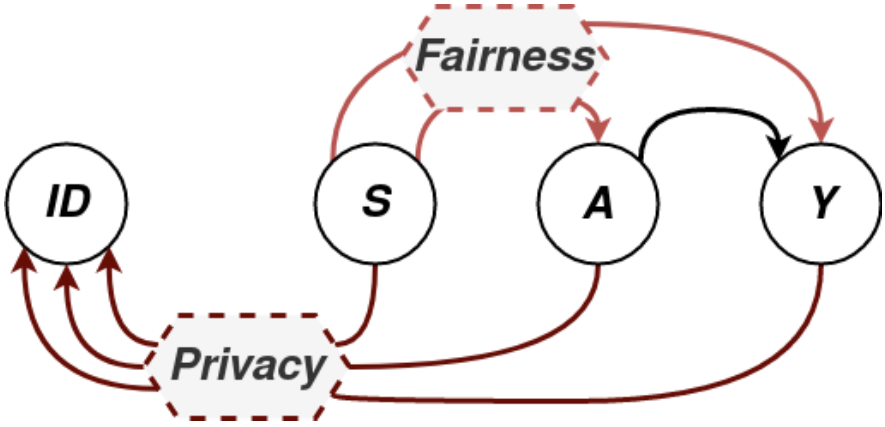


FIGURE 1.5 – L'anonymisation s'intéresse aux liens entre *ID* et les autres attributs, tandis que l'équité s'intéresse aux liens entre *S* et les autres attributs, démontrant là aussi les interactions potentielles entre les deux concepts.

TABLEAU 1.5 – Exemple de classes d'équivalence en considérant les valeurs des attributs *Pays d'origine* et *Niveau d'éducation*.

	Pays d'origine	*Niveau d'éducation*
C_1	Belgique	Maîtrise
C_2	Tchad	Doctorat

Il existe plusieurs modèles de respect de la vie privée tels que la *k-anonymité* (SWEENEY, 2002), la *t-proximité* (LI, LI et VENKATASUBRAMANIAN, 2006) et la *confidentialité différentielle (DP)* (DWORK, ROTH et al., 2014). Ces modèles peuvent s'appuyer notamment sur des classes d'équivalence, c'est-à-dire des ensembles de profils qui partagent les mêmes quasi-identifiants. Par exemple, en considérant les attributs *Pays d'origine* et *Niveau d'éducation* comme quasi-identifiants, on pourrait définir deux classes d'équivalence C_1 et C_2 tel que présentées dans le tableau 1.5. Le k anonymat demande à ce que la cardinalité de tous les ensembles de profils partageant des quasi-identifiants similaires soit supérieure à k. En reprenant l'exemple précédent, cette condition revient à ce que les classes C_1 et C_2 contiennent au moins k profils (on doit avoir au moins k profils de personnes venant de *Belgique* et ayant une *Maîtrise*, et au moins k profils de personnes venant du *Tchad* et détentrice d'un doctorat). La t-proximité requiert que la distribution d'un attribut sensible dans chaque classe d'équivalence, soit au plus éloignée d'une distance t de la distribution de ce même attribut dans l'ensemble de données. (RUGGIERI, 2014) a démontré qu'il existe une relation entre la t-proximité et l'équité de groupe, en montrant notamment que si la t-proximité est satisfaite, alors la différence des proportions de décisions négatives entre les groupes est bornée par un seuil proportionnel à t. Ainsi, il propose des adaptations des algorithmes d'anonymisation pour les appliquer dans le contexte de l'amélioration de l'équité des décisions. Ces adaptations consistent en l'imitation des algorithmes de généralisation pour l'anonymisation des données, sauf dans la partie où les valeurs de l'attribut sensible peuvent être généralisées (afin d'éviter de modifier les différents groupes).

DWORK, HARDT, PITASSI, REINGOLD et al. (2012) ont été parmi les premiers à établir un lien formel entre la confidentialité différentielle et l'équité. Plus précisément, ils ont démontré que l'équité individuelle assurée par la condition de Lipschitz (*cf.* section 1.4) est une généralisation de la confidentialité différentielle. En d'autres termes, un modèle M satisfait la ϵ-*DP* si et seulement si M satisfait la propriété (D_∞, d)-Lipschitz, avec $d = \epsilon|x \triangle z|$ où $|x \triangle z|$ représente le nombre de différences entre les ensembles de données x et z. Cette relation devient observable si l'on exprime dans

le formalisme de la $(\epsilon, 0) - DP$.

Définition 15 (Confidentialité différentielle (Dwork, Roth et al., 2014)). *Un mécanisme M défini de $\mathcal{X}$ vers $\mathcal{Y}$ est dit $(\epsilon, 0) - DP$ si pour tous profils x et $z \in \mathcal{X}$:*

$$P(M(x) \in \mathcal{Y}) \leq exp(\epsilon) P(M(z) \in \mathcal{Y}) \tag{1.23}$$

Autrement dit, la confidentialité est préservée si les distributions de sortie d'un algorithme sur deux ensembles de données voisins sont presque indiscernables (Chang et Shokri, 2021). L'équation 1.23 peut aussi être écrite sous la forme de l'équation 1.24.

$$\frac{P(M(x) \in \mathcal{Y})}{P(M(z) \in \mathcal{Y})} \leq exp(\epsilon) \tag{1.24}$$

On observe ainsi que l'équation 1.24 correspond à une variation de D_∞.

Chang et Shokri (2021) ont évalué l'impact d'une méthode d'équité sur l'inférence de l'appartenance d'un profil à l'ensemble de données qui a servi à entraîner un modèle. La vie privée peut être protégée dans ce cas si la présence ou l'absence d'un profil dans l'ensemble d'entraînement a une influence limitée sur la sortie de l'algorithme d'apprentissage, c'est-à-dire le modèle lui-même. Dans cette recherche, les auteurs ont montré qu'en modifiant un algorithme d'anonymisation pour satisfaire l'équité (Agarwal, Beygelzimer, Dudík, Langford et al., 2018), les bris de vie privée du groupe protégés seront plus importants. Plus précisément, il a été prouvé que l'application des contraintes d'équité force la mémorisation par le modèle des données du groupe protégé dont les décisions sont négatives. La mémorisation des données est plus prononcée lorsque la taille du groupe sensible est réduite. À l'inverse, Bagdasaryan, Poursaeed et Shmatikov (2019) ont montré que la DP appliquée durant l'entraînement d'un modèle induit une perte en exactitude de prédiction qui est plus importante dans le groupe sous-représenté. Cet écart de performance serait dû aux procédures d'arrondi du gradient et d'addition du bruit utilisés dans l'entraînement de modèles sous contrainte de la DP.

Jagielski, Kearns, Mao, Oprea et al. (2019) et Tran, Fioretto et Van Hentenryck (2021) proposent d'améliorer simultanément l'équité et la confidentialité différentielle par rapport à l'attribut sensible. Plus précisément, Jagielski, Kearns, Mao, Oprea et al. (2019) modifie l'approche de post-traitement (Hardt, Price, Srebro et al., 2016) utilisée pour améliorer l'égalité des chances,

en introduisant de l'aléa tirée d'une distribution laplacienne dans la variable représentant les vrais positifs et les faux négatifs. Tran, Fioretto et Van Hentenryck (2021) introduisent la DP sous forme d'un bruit calibré dans les gradients de l'optimisation.

De manière plus générale, Cummings, Gupta, Kimpara et Morgenstern (2019) ont démontré que la construction d'un classifieur équitable parfait (par rapport à l'égalité des chances et la α-discrimination) et qui satisferait la DP est impossible. En effet, si un tel algorithme existait, il générerait des modèles dont les prédictions sont équivalentes à des fonctions constantes. Ils présentent néanmoins un algorithme basé sur le mécanisme d'exponentiation (Dwork, McSherry, Nissim et Smith, 2006) qui permet de retourner un modèle de prédiction respectueux de la vie privée et satisfaisant une approximative des contraintes d'équité mesurée par l'égalité des chances.

Agarwal (2020) a étendu la démonstration de l'impossibilité à une plus large famille de métrique de l'équité de groupe et prouvé que cette impossibilité tient même pour des versions relâchées des métriques d'équité. La démonstration de cette impossibilité repose sur la définition de la confidentialité différentielle, qui requiert que la sortie d'un algorithme soit quasiment invariante pour deux bases de données (et par extension deux distributions) «voisines» Ainsi, s'il existe un algorithme permettant de construire un modèle de prédiction assurant la confidentialité différentielle et équitable pour une distribution $\mathcal{D}$, alors cet algorithme produirait le même modèle pour toutes les distributions voisines de $\mathcal{D}$. Par conséquent, si l'algorithme produit des modèles différents des fonctions constantes, alors on peut toujours trouver deux distributions voisines de $\mathcal{D}$ pour lesquelles le modèle n'est pas équitable. En effet, dans l'hypothèse où le modèle n'est pas constant, on peut considérer la distribution formée par l'ensemble des données du groupe protégé ayant reçu des décisions négatives prédites et l'ensemble des profils du groupe privilégié dont les décisions prédites sont positives. Ce modèle ne serait donc pas équitable pour ces distributions, ce qui contredirait l'hypothèse du voisinage de la confidentialité différentielle.

Considérant les attaques du type inférence d'attributs, Aalmoes, Duddu et Boutet (2022) ont proposé Diakaios, une méthode qui a pour but d'inférer la valeur de l'attribut sensible (binaire dans ce cas) des personnes à partir des prédictions d'un modèle entraîné dans l'objectif d'améliorer l'équité. L'approche proposée utilise des données auxiliaires (essentiellement des données provenant de la même distribution que les données d'entraînement du modèle attaqué) pour construire un modèle d'attaque utilisant les prédictions du modèle attaqué pour inférer l'attribut sensible. Elle optimise ensuite une fonction permettant d'obtenir le seuil de probabilité optimal de l'adversaire

pour décider du groupe d'appartenance de chaque personne. Le succès de leur approche montre ainsi que les modèles d'équité peuvent introduire des bris de vie privée.

À partir de l'ensemble de ces recherches susmentionnées, on peut constater que la protection de la vie privée et l'équité ont une relation intriquée où l'une des propriétés à un impact non négligeable sur l'autre.

1.9 Conclusions de notre étude générale

Dans ce chapitre, nous avons résumé certaines directions de recherche explorées actuellement dans le contexte de l'équité. Cependant, cette étude est nécessairement partielle et il existe de nombreux articles de recherches généraux (*surveys* ou état de l'art) qui étudient en profondeur différents aspects de l'équité et qui explorent ces différentes notions dans divers domaines (Barocas, Hardt et Narayanan, 2019 ; Carey et Wu, 2022 ; Caton et Haas, 2020 ; Fabris, Messina, Silvello et Susto, 2022 ; Jones, Hickey, Di Stefano, Dhanjal et al., 2020 ; Mehrabi, Morstatter, Saxena, Lerman et al., 2021 ; Mitchell, Potash, Barocas, D'Amour et al., 2018 ; Romei et Ruggieri, 2014 ; Ruggieri, Hajian, Kamiran et Zhang, 2014 ; Soremekun, Papadakis, Cordy et Traon, 2022 ; Tian, Zhu, Liu et Zhou, 2022 ; Verma et Rubin, 2018 ; Zehlike, Yang et Stoyanovich, 2021). Nous n'avons pas non plus abordé les aspects de l'équité dans plusieurs autres domaines, tels que le classement (*ranking*), les systèmes de recommandations, les données non structurées (images, audio, etc.) ou encore les formes d'apprentissage autres que l'apprentissage supervisé. Nous référons le lecteur intéressé aux articles de Zehlike, Yang et Stoyanovich (2021), qui étudient l'équité dans le contexte des algorithmes de classement (*ranking*), et de Tian, Zhu, Liu et Zhou (2022) qui l'évaluent pour les images.

De l'étude que nous avons menée, il ressort qu'il existe une pléthore d'approches ayant pour objectif d'améliorer l'équité. En effet, cette dernière peut être définie de plusieurs façons différentes, et pour chacune de ces définitions, des familles d'approches y sont associées. L'équité de groupe est la plus étudiée, due à la simplicité d'implémentation et d'intégration des métriques au sein des modèles d'apprentissage déjà existants. La grande quantité et diversité des approches rend la compréhension difficile pour les personnes qui découvrent le domaine de l'équité.

De plus, la discordance entre les métriques de quantification et la grande variété des techniques possibles rendent le choix de la technique à utiliser très difficile et peuvent conduire à des situations

difficiles dans lesquelles des personnes peuvent se sentir discriminées, bien que le modèle ait été optimisé pour satisfaire une définition différente de l'équité. Très peu de recherches s'intéressent à la standardisation et à la comparaison des approches dans différents contextes ainsi que leur robustesse. Une étude récente (Jones, Hickey, Di Stefano, Dhanjal et al., 2020) propose un premier formalisme et développe un processus standard de comparaison des approches.

Un autre aspect très peu étudié dans la littérature sur les méthodes d'amélioration de l'équité est l'impact à long terme des méthodes d'amélioration de l'équité. En effet, la plupart des méthodes proposées supposent un environnement statique, qui n'évoluerait pas dans le temps. Ainsi, non seulement les modèles appris peuvent devenir désuets à cause des variations des distributions des données, ils peuvent aussi être la cause de ces variations. En effet, l'objectif des méthodes étant de la prise de décision équitable, ces modèles peuvent promouvoir certains groupes qui ne l'auraient été sans l'ajout des contraintes d'équité. Cependant, la difficulté dans l'analyse de l'impact à long terme reste la possibilité de simulation d'un état du monde que l'on suppose qui serait le résultat de l'application des méthodes d'amélioration de l'équité des décisions.

Enfin, la recherche en équité dans l'apprentissage automatique souffre d'un manque criant de recul, car le domaine est très axé sur les problématiques mathématiques et commence seulement à intégrer les remarques et différents retours des études menées par les sciences sociales. Un défi à long terme serait par conséquent de réussir à combiner les recherches en sciences sociales et le domaine très formel des sciences mathématiques appliquées à l'apprentissage automatique.

Résumé et chapitres suivants. Nous avons introduit tout au long de ce chapitre l'origine des biais, présenté les définitions de l'équité des décisions ainsi que les métriques pour la quantifier, ensuite, nous avons présenté les mécanismes d'amélioration de l'équité. Les recherches que nous avons menées durant cette thèse portent principalement sur les biais existant dans les données d'entrainement, avec un accent mis sur la discrimination indirecte. L'objectif que nous souhaitions atteindre est l'amélioration de l'équité des décisions par la prévention de l'inférence de l'attribut sensible. Les approches que nous avons développées sont principalement des approches de prétraitement des données, car celles-ci permettent l'utilisation de l'ensemble de données prétraitées pour plusieurs tâches d'analyses, en plus d'offrir la possibilité de publication des données. Ainsi, les données peuvent être utilisées par différents organismes tout en maintenant un niveau élevé de protection des informations sensibles. Puisque la prévention d'inférence de l'attribut sensible peut-être reliée à l'équité de groupes, nous

quantifions aussi cette équité, pour caractériser le comportement de nos approches.

CHAPITRE II

ASSAINISSEMENT LOCAL DES DONNÉES POUR L'AMÉLIORATION DE L'ÉQUITÉ PAR UN ENTRAÎNEMENT ANTAGONISTE

Notre première direction de recherche a été menée au travers du projet *Local Data Debiasing for Fairness Based on Generative Adversarial Training* (connu sous le sigle de *GANSan*), publié en 2021 dans le numéro spécial *Interpretability, Accountability and Robustness in Machine Learning* du journal *Algorithms* (Ulrich Aïvodji, François Bidet, Sébastien Gambs, Rosin Claude Ngueveu et al. (mars 2021). « Local Data Debiasing for Fairness Based on Generative Adversarial Training ». In : *Algorithms* 14.3, p. 87. issn : 1999-4893). Je suis l'auteur principal de cette recherche ainsi que le concepteur principal de la méthode. J'ai aussi écrit les codes associés, analysé les résultats et participé activement à la rédaction de l'article de recherche.

GANSan est une approche de prétraitement des données dont l'objectif est de réduire le biais mesuré dans le jeu de données au travers d'un attribut que l'on juge sensible, tout en introduisant le moins de modifications possibles. Ainsi, les corrélations utiles peuvent être maintenues et le jeu de données peut être utilisé pour plusieurs tâches subséquentes. Cette approche présente aussi l'avantage de préserver l'interprétabilité des données en ne changeant pas d'espace de représentation, et permet à toute personne de transformer son profil par un assainissement local (sur son téléphone par exemple) avant de publier une version assainie de ce profil. Grâce à ces propriétés, notre approche se distingue ainsi des méthodes de l'état de l'art.

De manière générale, notre approche GANSan s'appuie sur le principe de l'équité par la prévention d'inférence de l'attribut sensible. Pour rappel, l'équité par la prévention d'inférence de l'attribut sensible consiste en la suppression de corrélations entre cet attribut sensible et les autres attributs du jeu de données, permettant à un processus de décision utilisant les données modifiées de rendre une décision sans aucune information concernant cet élément sensible (aucune influence de l'attribut

sensible). La discrimination (directe ou indirecte) est empêchée puisque la caractéristique permettant d'attribuer des traitements différents est inconnue.

Ainsi, GANSan transforme les données en y introduisant un minimum de perturbations nécessaires, de sorte que l'inférence de l'attribut sensible, telle que mesurée à travers différents mécanismes de classification, soit difficile.

2.1 GANSan : cadriciel

GANSan est inspiré des RAG (section 1.7). De ce fait, il est aussi composé de deux modèles, un générateur et un discriminateur, tout comme ces réseaux adversariaux. Cependant, à la différence des RAG qui génèrent des données à partir du bruit X, le générateur, que nous avons nommé *assainisseur* dans notre approche, reçoit en intrant des données réelles, et produit en extrant des données ayant la même représentation que les données en intrants. Dans notre approche, le discriminateur est utilisé pour déduire l'attribut sensible S à partir des données produites, l'incapacité du discriminateur à prédire S constituent le critère de succès de GANSan.

Formellement, étant donné un ensemble de données R, l'objectif de GANSan est de construire une fonction S_{an} (appelée *assainisseur*) qui modifie les données R tout en satisfaisant du mieux possible deux critères :

- La distance entre les données originales R et les données assainies $\bar{R} = S_{an}(R) = \{\bar{A}, \bar{Y}\}$ doit être minimale. $\bar{A}$ représente la version assainie de A, $\bar{Y}$ correspond à la version assainie de Y.
- L'attribut sensible S ne doit pas pouvoir être prédit à partir de $\bar{R}$ (ou avec une faible précision). Idéalement, cela se traduirait par l'impossibilité de trouver une fonction f telle que $S = f(\bar{R})$.

Comme mentionné précédemment, le discriminateur dans GANSan a pour objectif de construire la frontière de décision permettant de distinguer les groupes sensibles définis par l'attribut binaire S, au lieu de prédire si les données modifiées proviennent bel et bien de la distribution originale comme c'est le cas dans les RAG et les RAG conditionnels.

L'assainisseur se comporte de ce fait comme un auto-encodeur (Rumelhart, Hinton et Williams, 1985) dont la tâche est de reconstruire les données originales tout en s'assurant que le discriminateur D ne puisse pas inférer S. La figure 2.1 présente une vue d'ensemble de l'entraînement de l'assainisseur tandis que nous présentons dans l'algorithme 1 les détails de la procédure d'entraînement.

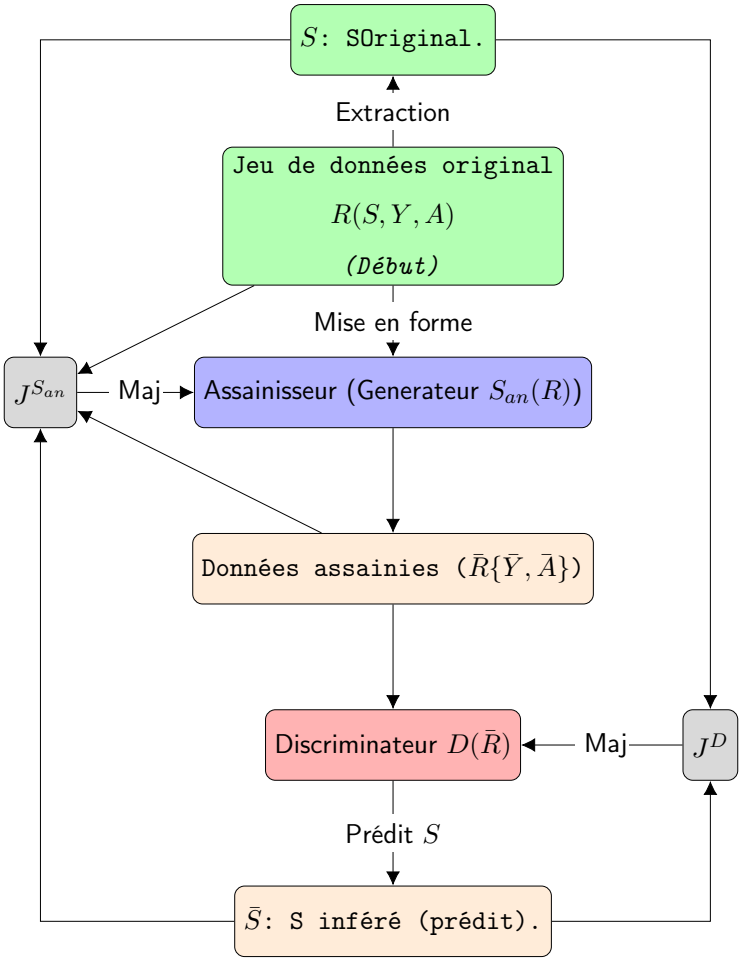


FIGURE 2.1 – Vue d'ensemble de l'approche GANSan. L'objectif du discriminateur est la prédiction de S à partir des extrants de l'assainisseur, $\bar{R}$. L'approche a pour objectif de minimiser les fonctions de perte du discriminateur et de l'assainisseur, qui sont respectivement J^D et $J^{S_{an}}$.

La première étape correspond à l'entraînement de l'assainisseur S_{an} (algorithme 1, lignes 7-17), qui apprend la distribution empirique de l'attribut sensible afin d'introduire le minimum de perturbations pour tromper le discriminateur dans sa prédiction de S. Pour cela, S_{an} utilise l'ensemble de données original R (composé de A, Y et de S) ainsi que de valeurs aléatoires X, et produit une version assainie $\bar{R} = \{\bar{A}, \bar{Y}\}$ qui ne contient pas d'information concernant S. Plus précisément, l'assainisseur prend en entrée un vecteur v_i composé du profil r_i de la personne i et d'une valeur aléatoire x_i ($v = \{r_i, x_i\} = \{a_i, y_i, x_i\}$), et produit un nouveau profil $\bar{r}_i$ pour cette même personne, $\bar{r}_i = \{a_i, y_i\}$, tel que D ne serait pas en mesure de déterminer correctement la valeur de s_i de chaque profil r_i.

Algorithm 1 Procédure d'entraînement de GANSan

Intrants : $D = \{A, Y, S\}$, MaxEpochs, d_{iter}, batchSize, α
Extrants : S_{an}, D

▷ Initialisation

S_{an}, D, $R_{isc} = \text{mélanger}(R)$
$\text{Iterations} = \frac{|D|}{\text{batchSize}}$
for e $\in \{1, .., \text{MaxEpochs}\}$ **do**
    for i $\in \{1, .., \text{Iterations}\}$ **do**
        Prélever un lot de donnée B de taille batchSize à partir de R
        S_B : Extraire S de B
        $\{\bar{A}, \bar{Y}\} = S_{an}(B)$
        $e_{A_i} = \frac{1}{batchSize} \cdot \sum_{n=1}^{\text{batchSize}} |A_i^n - \bar{A}_i^n|$

▷ Calculer le vecteur d'erreur de reconstruction

        $\vec{J}^{San} = (1 - \alpha) \cdot (e_{A_1}, e_{A_2}, e_{A_3}, ..., e_{A_d}, e_Y)^T$

▷ Calculer l'erreur sur l'attribut sensible

        $d_S = \alpha * (\frac{1}{2} - \text{BER}(D(S_{an}(B), S_B))$

▷ Concaténer toutes les erreurs calculées

        $\vec{J}^{San} = \text{concat}(\vec{J}^{San}, d_S)$
        for $loss \in \vec{J}^{San}$ **do**
            Rétro propagation de $loss$
            Mise à jour des poids de S_{an}
        end for
        for l $\in \{1, .., d_{iter}\}$ **do**
            Extraire le lot B de taille $batchSize$ de R_{isc}
            S_B : Extraire S de B
            $d_{disc} = \text{MSE}(S_B, D(S_{an}(B)))$
            Rétro propagation de $Loss$
            Mise à jour des poids de D_{isc}
        end for
    end for
    Sauvegarder les états des modèles S_{an} et D
end for

Les valeurs aléatoires X sont tirées d'une distribution uniforme entre 0 et 1 et elles permettent d'éviter le surapprentissage et la mémorisation, rendant ainsi plus difficile la transformation inverse des données assainies vers les données originales (l'assainissement est rendu probabiliste). En consé-

quence, plusieurs versions assainies d'un même profil peuvent être générées par l'assainissement, chacune correspondant à l'assainissement du profil original couplé à une valeur tirée de X. De même, l'attribut sensible utilisé en entrée de l'assainisseur permet d'orienter la transformation des données en indiquant le groupe d'appartenance de chaque personne. Ainsi, les données du groupe privilégié pourraient être transformées vers le groupe protégé, et celles du groupe protégé pourraient être ramenées près de la distribution du groupe privilégié. L'utilisation de l'attribut sensible et du bruit offrent aussi une protection supplémentaire contre les attaques d'appartenance (Salem, Zhang, Humbert, Berrang et al., 2018 ; Shokri, Stronati, Song et Shmatikov, 2017 ; Song, Shokri et Mittal, 2019 ; Zhang, Yu, Sun, Li et al., 2020) (*membership attacks*), dont l'objectif est d'identifier les données ayant été utilisées dans l'ensemble d'entraînement d'un modèle.

La seconde étape consiste en l'entraînement du discriminateur D pour la prédiction de l'attribut sensible à partir de la sortie de l'assainisseur $S_{an}(B)$ (algorithme 1, Lignes 18-24), avec B un sous-ensemble de R utilisé durant l'entraînement. Le succès du discriminateur indique une mauvaise capacité de l'assainisseur à protéger l'attribut sensible, et par conséquent un risque plus élevé de discrimination. Ces deux étapes entre l'assainisseur et le discriminateur s'exécutent itérativement jusqu'à la convergence.

2.1.1 Entraînement de GANSan

Soit $\bar{S}$ la prédiction de S par le discriminateur D à partir des données produites par l'assainisseur $S_{an}(R)$. L'objectif du discriminateur est d'être le plus performant possible dans ses prédictions $\bar{S}$. Pour cela, D minimise la moyenne quadratique des erreurs de prédictions *MSE* (*Mean Squarred Error* en anglais) présentée dans l'équation suivante.

$$MSE(S, \bar{S}) = \frac{1}{N} \sum_{i=1}^{N} (s_i - \bar{s}_i)^2. \tag{2.1}$$

La fonction objective du discriminateur est de ce fait : $J^D(S, \bar{S}) = MSE(S, \bar{S})$ alors quela fonction objective de l'assainisseur $J^{S_{an}}(.,.,.)$ est présentée en équation 2.2 :

$$J^{S_{an}}(R,\ S_{an},\ D) = (1 - \alpha) \times d_r(R,\ S_{an}(R)) + \alpha \times d_s(S, \bar{S}) \tag{2.2}$$

Celle-ci est composée de fonctions $d_r(.,.)$ et $d_s(.,.)$, qui sont respectivement liées à nos différentes contraintes à savoir la minimisation des modifications introduites dans les données et la protection de l'information sensible.

Pour la minimisation des modifications introduites dans les données, nous utilisons l'erreur absolue moyenne (*Mean Absolute Error (MAE)*, equation 2.3) pour instancier $d_r(.,.)$.

$$d_r(R, \bar{R}) = MAE(R, \bar{R}) = \frac{1}{N} \frac{1}{d-1} \sum_{i=1}^{N} \left[\sum_{j=1}^{d-2} \mid a_{i,j} - \bar{a}_{i,j} \mid + \mid y_i - \bar{y}_i \mid \right] \tag{2.3}$$

avec $a_{i,j}$ le $j^{\text{ême}}$ attribut non sensible du profil i. Nous avons aussi utilisé la fonction MSE pour d_r, cependant, nos expériences préliminaires ont montrées que MAE est plus adaptée que MSE dans la minimisation des pertes d'information.

La protection de l'attribut sensible se fait par l'intermédiaire de la fonction d_s, qui mesure à l'aide de D la quantité d'information concernant S contenue dans les données modifiées $\bar{R}$. L'objectif de l'assainissement étant de minimiser la quantité d'information sensible, celle-ci se fait au travers de l'incapacité de D à prédire S à partir des attributs assainis $\{\bar{A}, \bar{Y}\}$. Ceci se traduit par la maximisation de l'erreur de prédiction de D, qui correspond à la maximisation du BER (le maximum, dans le cas binaire est de 1/2). On peut ainsi introduire directement le BER (équation 2.4) :

$$\begin{aligned} d_s &= \frac{1}{2} - BER(D(\bar{A}, \bar{Y}), S) \\ &= \frac{1}{2} - BER(\bar{S}, S) \end{aligned} \tag{2.4}$$

Enfin, dans la fonction objective $J^{S_{an}}$, le paramètre α contrôle l'importance relative de la protection de l'attribut sensible par rapport à la minimisation des modifications introduites. Plus précisément, α varie entre 0 et 1 : lorsque α tend vers 0, la minimisation des modifications est priorisée par rapport à la protection de l'attribut sensible. Inversement, avec $\alpha \approx 1$, l'assainisseur priorisera la protection de l'attribut sensible au détriment de la qualité des données produites en sortie du modèle. La valeur optimale de α dépend du niveau de protection souhaitée par la personne utilsatrice, de l'ensemble de données et de la structure des modèles. Ainsi, pour des valeurs de α proches ou identiques, les tendances obtenues pourraient être significativement différentes d'un jeu de données à un autre.

Les RAG sont sujets au phénomène de *mode collapse* qui correspond au fait que le générateur n'arrive pas à modéliser la diversité des données et génère des échantillons qui sont très similaires, sinon identiques (Srivastava, Valkov, Russell, Gutmann et al., 2017). Dans notre cas, ce phénomène correspond au fait que l'assainissement conduit à la transformation de tous les profils des personnes vers un profil unique. Pour éviter ce problème, nous avons décidé de calculer l'erreur absolue moyenne dans la fonction (d_r) sur chacun des attributs de l'ensemble de données (équation 2.5), sans calculer

la moyenne de ces erreurs sur tous ces attributs.

$$
\begin{aligned}
d_r(R,\bar{R}) = MAE(R,\bar{R}) &= \left\{MAE(A_1,\bar{A}_1),\ldots,MAE(A_{d-2},\bar{A}_{d-2}),MAE(Y,\bar{Y})\right\} \\
&= \left\{\left[\frac{1}{N}\sum_{i=1}^{N} \mid a_{i,1}-\bar{a}_{i,1} \mid\right],\ldots,\left[\frac{1}{N}\sum_{i=1}^{N} \mid a_{i,d-2}-\bar{a}_{i,d-2} \mid\right],\left[\frac{1}{N}\sum_{i=1}^{N} \mid y_i-\bar{y}_i \mid\right]\right\}
\end{aligned} \tag{2.5}
$$

Ainsi, la fonction objective de l'assainisseur s'écrit sous la forme présentée en équation 2.6 :

$$
J^{S_{an}}(R,\ S_{an},\ D) = \left\{ \begin{array}{c} (1-\alpha)\times MAE(A_1,\bar{A}_1),\ldots,(1-\alpha)\times MAE(A_{d-2},\bar{A}_{d-2}) \\ (1-\alpha)\times MAE(Y,\bar{Y}), \alpha\times(\frac{1}{2}-BER(D(\bar{A},\bar{Y}),S)) \end{array} \right\} \tag{2.6}
$$

Pour chaque itération de l'entraînement de S_{an}, $J^{S_{an}}$ retourne un ensemble d'erreurs calculées sur chacun des attributs de l'ensemble de données. La descente de gradient se fait en itérant sur chacun des éléments retourné par $J^{S_{an}}$. L'objectif de cette modification est de pouvoir introduire un peu d'aléa dans les gradients calculés, de sorte que l'optimisation par descente du gradient ne restent pas bloquée dans un minimum local.

Finalement, l'utilisation de GANSan peut se faire en prenant en compte ou non la décision. Dans l'entraînement de notre modèle, nous considérons que la décision est biaisée et fait partie des attributs à assainir en introduisant le terme $MAE(Y,\bar{Y})$ au sein de la fonction d_r. Toutefois, l'assainissement peut aussi se faire en excluant la décision des attributs à assainir (dans le cas où la décision serait une information critique à ne pas modifier). L'exclusion de la décision se fait en excluant le terme $MAE(Y,\bar{Y})$ de la fonction de reconstruction d_r. Le code de notre méthode est disponible sur le dépôt suivant : `https://gitlab.privsec.ca/Rosin/gansan/`

2.2 GANSan : cadre expérimental

Ci-après, nous présenterons le cadre expérimental utilisé pour mesurer les performances de GANSan.

2.2.1 Mesures de performance

Nous évaluerons les performances de GANSan suivant l'amélioration de l'équité et l'utilité préservée dans les données. GANSan étant utilisé pour protéger l'attribut sensible, cet objectif sera mesuré par l'incapacité d'un prédicteur *Adv*, que nous appellerons *adversaire*, à prédire l'attribut sensible. Nous quantifierons cela notamment au travers du *Taux d'erreur pondéré (BER)* (Feldman, Friedler, Moeller, Scheidegger et al., 2015) et de l'*exactitude de prédiction (SAcc)* (*cf.* chapitre 1, section 1.4). Nous mesurerons aussi la parité démographique (DemoParity, équation (1.5)) et l'égalité des chances (EqGap, équation (1.8)) qui permettront aussi d'évaluer l'amélioration de l'équité.

Concernant l'utilité, nous utiliserons la fidélité Fid pour représenter la préservation des données originales. Fid est calculée en mesurant la distance entre les données originales et leur version assainies. La distance sera instanciée par la norme L_2, étant donné que celle-ci n'avantage pas un ou plusieurs attributs par rapport aux autres.

Définition 16 (Fidélité)**.** *Étant donné les attributs non sensibles A, de décision Y et leurs transformations respectives $\bar{A}$ et $\bar{Y}$, la fidélité Fid mesure sur l'ensemble de données la proximité des transformations de leurs valeurs originales respectives :*

$$Fid = 1 - \frac{1}{N}(L_2(\{\bar{A}, \bar{Y}\}, \{A, Y\}))^2 = 1 - \frac{1}{N}\sum_{i=1}^{N}[\frac{\sum_{j=1}^{d-2}(a_i - \bar{a}_i)^2}{d-2} + (y_i - \bar{y}_i)^2]. \tag{2.7}$$

Le dénominateur dans l'équation 4.13 correspond au fait que nous ne considérons qu'un seul attribut sensible, de ce fait, l'ensemble de données R est composé de $d-2$ attributs qui ne sont ni des attributs sensibles, ni la décision Y. En ce qui concerne la notion de distance, notre approche ne nécessite aucune hypothèse ou pré-requis sur celle utilisée, bien que certaines mesures de distance puissent être meilleures que d'autres en fonction du type de données. Par exemple, on pourrait considérer une notion de distance qui serait conditionnelle aux valeurs des attributs, pour tenir compte de certaines valeurs très peu représentées dans l'ensemble de données, ou tout simplement de leur signification (par exemple, des attributs qui représenteraient le degré de crime).

L'objectif étant de maintenir la plus grande fidélité possible entre les données originales et leur version assainies, l'assainisseur doit produire des données ayant la valeur de Fid la plus haute possible (c'est-à-dire la plus proche possible de 1) pour une protection donnée. Cependant, il existe des situations dans lesquelles la fidélité est insuffisante pour quantifier l'utilité du jeu de données. Par exemple, pour un ensemble de données composé de profils proches, la transformation des profils vers un unique profil médian produirait une fidélité élevée (du fait de la proximité des profils) bien que le jeu de données soit inutilisable.

Ainsi, pour quantifier la capacité de l'assainisseur à préserver la diversité et la représentativité des données, nous avons introduit la diversité $Diversité$ définie de la manière suivante :

Définition 17 (Diversité)**.** *Soit un ensemble de données composé de N profils $r_i (i \in \{1, \ldots, N)$ et de dimension d. La mesure diversité quantifie la moyenne des distances entre les différentes personnes*

d'un même ensemble de données :

$$Diversité = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{N-1}\sum_{j=1,j\neq i}^{N}\sqrt{\frac{1}{d}\sum_{l=1}^{d}(\bar{r}_{i,l}-\bar{r}_{j,l})^2}\right] \tag{2.8}$$

$\bar{r}_{i,k}$ représente le k^{ime} attribut de la version assainie du profil r_i.

Diversité varie dans l'intervalle $[0,\infty[$, avec 0 indiquant l'identité de tous les profils de l'ensemble de données, et ∞ indiquant les variations non bornées et la grande hétérogénéité des données.

La fidélité et la diversité sont complémentaires dans l'analyse du jeu de données assainies $\bar{R}$. En effet, la première mesure la quantité de dommage introduite par l'assainissement tandis que la seconde mesure la préservation de l'espace des possibles valeurs du jeu de données. Cependant, ces métriques ne permettent pas d'obtenir une compréhension qualitative des modifications. En d'autres termes, elles ne permettent pas à une personne de se représenter les modifications introduites qualitativement dans le jeu de données pour différentes valeurs de fidélité. Pour cela, nous apporterons aussi dans notre recherche une discussion qualitative sur les modifications qu'impliquerait une valeur donnée de fidélité et de protection de l'attribut sensible.

Enfin, nous évaluerons la perte d'utilité (pour des tâches subséquentes) induite par l'assainissement, en mesurant l'exactitude de prédiction de la décision Acc_Y. Plus précisément, la différence de valeur de Acc_Y entre un classifieur entraîné sur les données originales et un autre entraîné sur les données assainies peut être utilisée comme mesure de perte d'utilité (pour la tâche en question) due à l'assainissement.

2.2.2 Description des ensembles de données

Nous avons évalué notre approche sur deux ensembles de données communément utilisés dans la littérature sur l'équité, *Adult Census Income* et *German Credit*. Ces ensembles de données présentés de manière détaillée dans le chapitre 1 en section 1.5, nous ferons simplement un bref rappel dans cette section. L'ensemble *Adult Census Income* est composé de $45,222$ profils qui comportent chacun 15 attributs. L'attribut sensible utilisé pour cet ensemble est le *genre* (*gender*), qui est homme ou femme, tandis que l'attribut de décision est le niveau de revenu (décision positive pour des revenus supérieurs à $50k\$$). *German Credit* est composé de 1000 profils de candidats pour un crédit, chacun étant décrit par 21 caractéristiques bancaires. L'attribut sensible est l'*âge*, transformé en attribut binaire discret en considérant les catégories $\hat{a}ge > 25$ et $\hat{a}ge <= 25$ ans. Selon Kamiran et

Calders (2009), cette discrétisation maximiserait la discrimination observée dans l'ensemble de données. L'attribut de décision dans cet ensemble est la qualité du client (bon ou mauvais payeur). Rappelons que le tableau 1.3 décrit les distributions de ces ensembles de données.

Prétraitement des ensembles de données. L'étape de prétraitement consiste à formater les données de manière à ce qu'elles puissent être utilisées par les modèles de réseaux neuronaux. La première étape consiste à encoder les attributs catégoriels et les attributs numériques dont le domaine est discret et composé de moins de cinq de valeurs, puis à les mettre à l'échelle entre 0 et 1. L'encodage utilisé est le *one-hot encoding* qui transforme les attributs catégoriques en une succession d'attributs numériques qui prennent la valeur 1 à la position originale et 0 dans le cas contraire. Un exemple est présenté dans le tableau 2.1, où l'attribut *métier* est transformé en *métier=pompier* et *métier=policier*.

Tableau 2.1 – Transformation de l'attribut catégorique original *métier* en une succession d'attributs numériques. Le premier profil (r_i) prend la valeur 1 sur la colonne *métier=pompier* et 0 sur la colonne *métier=policier* étant donné que le métier original de celui-ci est celui de pompier. Le même processus est appliqué aux attributs numériques discrets ayant moins de cinq valeurs.

	Métier	*Métier = Pompier*	*Métier = Policier*
r_i	*Pompier*	1	0
r_j	*Policier*	0	1

En outre, dans le cas de l'ensemble de données Adult, nous devons appliquer un logarithme aux valeurs des attributs *capital-gain* et *capital-loss* avant toute autre étape de mise en forme des données, ceci afin de limiter la dispersion des valeurs de ces attributs. En figure 2.2, nous présentons les valeurs originales de ces deux attributs, et celles obtenues après application du logarithme (log-capital-gain = $log(1 + \text{capital-gain})$ [1]). La grande dispersion des valeurs et la prédominance de la valeur 0 ferait en sorte que les modèles d'assainissement ne seraient pas en mesure d'apprendre facilement la distribution des valeurs de cet attribut. Après l'application du logarithme, l'intervalle des valeurs passe de $[0, 10^5]$ à $[0, 12]$.

1. On ajoute 1 à toutes les valeurs de ces attributs pour éviter $log(0)$.

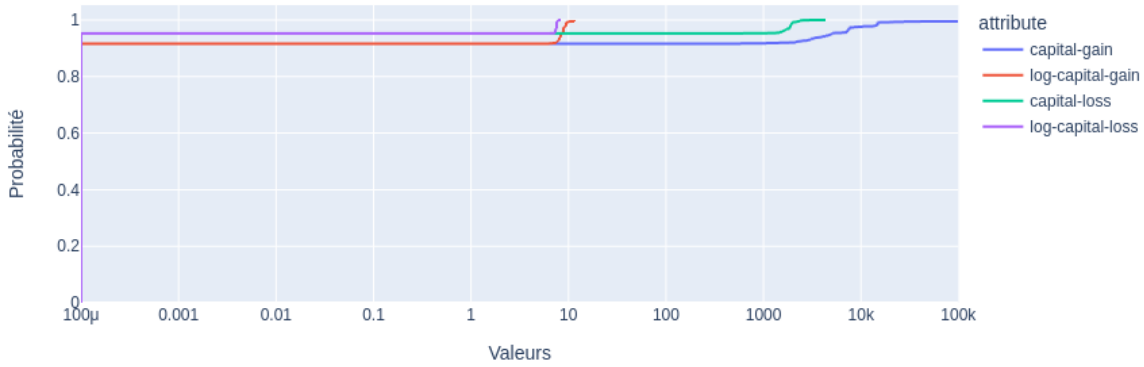


FIGURE 2.2 – Distributions des attributs capital-gain et capital-loss.

À la fin de l'assainissement, des étapes de post-traitement consistant à inverser les étapes de pré-traitement sont effectuées afin de redonner aux données générées leur forme originale.

2.2.3 Hyperparamètres des modèles

Nous présentons dans le tableau 2.2 la structure des réseaux de neurones ayant produit les meilleurs résultats sur les ensembles de données Adult et German Credit. La variable *taux d'apprentissage* représente le nombre de mises à jour du modèle concerné par itération d'entraînement. Par exemple, pour des tailles de sous-ensembles d'entraînement du discriminateur et de l'assainisseur égales à 64, à chaque itération i, les paramètres du discriminateur sont mis à jour 50 fois, ce qui signifie que le discriminateur est ainsi entraîné avec $64 \times 50 = 3200$ profils de personnes, pendant que l'assainisseur est entraîné avec uniquement $64 \times 1 = 64$ profils différents.

Le nombre total d'itérations est défini par le ratio $iterations = N/Bs$, avec N la taille de l'ensemble de données et Bs la taille du sous-ensemble d'entraînement. Nous avons choisi d'utiliser $ReLU$ et $LeakyReLU$ dans notre architecture car ces fonctions d'activation ont des performances supérieures à *Sigmoid* et *Tanh* (KRIZHEVSKY, SUTSKEVER et HINTON, 2012). De plus, $ReLU$ et $LeakyReLU$ ne sont pas sujettes à la saturation [2] (et la disparition [3]) des gradients, et produisent des représentations

2. Par exemple, pour la fonction sigmoid ($sig(x)$), des valeurs très grandes de x font en sorte que sa dérivée $\frac{\partial sig(x)}{\partial x}$ est très proche de 0, ce qui ne permet pas d'entraîner le réseau de neurones.

3. La disparition du gradient survient lorsque l'erreur rétro-propagée devient 0 et les poids du réseaux ne sont pas

Tableau 2.2 – Hyperparamètres des réseaux de neurones pour les ensembles Adult et German.

	Assainisseur	**Discriminateur**
Nombre de couches	3 × Linéaires	5 × Linéaires
Vitesse d'apprentissage (LR)	2×10^{-4}	2×10^{-4}
Fonction d'activation des couches cachées	ReLU	ReLU
Fonction d'activation des couches de sorties	LeakyReLU	LeakyReLU
Fonction de pertes	VectorLoss	MSE
Taux d'apprentissage	1	50
Taille des sous-ensembles d'entraînement (Batch size Bs)	64	64
Optimiseur	Adam	Adam

avec très peu de coefficients (Glorot, Bordes et Bengio, 2011). Concernant le paramètre de compromis α, nous avons choisi la progression $\alpha_i = 0.2 + 0.4\frac{2^i - 1}{2^{i-1}} avec\ i \in \{1, \ldots, 10\}$, afin de construire le front de Pareto, et de donner plus d'importance à l'équité dans la fonction objective, étant donné que nos expériences préliminaires ont montré la tendance de notre approche à optimiser plus facilement la reconstruction des données. Chacune de nos expériences a été menée pour un total de 40 époques. Une comparaison du temps d'exécution de GANSan comparativement à d'autres approches sera présentée dans la section 2.5.

2.2.4 Procédure d'entraînement

GANSan est évalué suivant l'utilité préservée, la protection de l'attribut sensible S et l'amélioration de l'équité, capturée respectivement par Fid, BER et $DemoParity$ (section 2.2.1). Pour mesurer les performances de notre approche, nous avons réalisé une validation croisée de 10 partitions. Ainsi, l'ensemble de données est divisé aléatoirement en 10 partitions. Pour chaque partition de la validation croisée, huit sous-ensembles sont utilisés pour l'entraînement, et les deux sous-ensembles restants sont utilisés respectivement pour la validation et le test. Le BER et le $SAcc$ sont calculés en utilisant le discriminateur interne de GANSan D, ainsi que trois classifieurs externes indépendants de GANSan : les *Machines à Vecteurs de Support* (*Support Vector Machines* ou *SVM*) (Cortes et Vapnik, 1995), un *Perceptron Multicouche* (*Multilayer Perceptron* ou *MLP*) (Popescu, Balas, Perescu-Popescu

mis à jour.

et Mastorakis, 2009) et *Gradient Boosting* (GB) (Friedman, 2002).

La plupart des approches de la littérature (cf. section 1.6.2) limitent la validation de leur approche à l'utilisation de classifieurs intégrés dans l'architecture de leur approche (autrement dit, aucune validation indépendante n'est effectuée). Dans le cadre de GANSan cela correspondrait à l'utilisation unique du discriminateur interne pour reporter les résultats obtenus. Or, les résultats ainsi présentés pourraient être biaisés du fait que la protection de l'attribut sensible n'est pas mesurée par des méthodes n'ayant jamais eu connaissance des données originales, qui pourraient exploiter des corrélations différentes et ainsi obtenir une meilleure qualité d'inférence de S. GANSan est validée en utilisant des classifieurs externes indépendants de notre système, qui sont de trois familles différentes (permettant de mesurer la protection sous différents mécanismes). Toutefois, nous reconnaissons que cette procédure de validation de résultat n'est pas parfaite. En effet, le fait qu'aucun classifieur utilisé n'arrive à inférer l'attribut sensible ne garantit pas qu'il n'existe aucune autre technique qui puisse obtenir de meilleures performances.

La sélection des meilleurs hyperparamètres est réalisée avec la procédure suivante : pour chacun des blocs et chacune des valeurs de α, l'assainisseur est entraîné pour une durée de 40 époques. À la fin de chaque époque, une version assainie de l'ensemble de données est générée à partir de l'état sauvegardé de l'assainisseur et les métriques associées BER, $SAcc$ et Fid sont calculées sur l'ensemble de validation et enregistrées. Parmi les 40 versions assainies du jeu de données de validation, nous choisissons celle dont les métriques sont les plus proches du point optimal $\{BER = 0,5, Fid = 1\}$ à partir de l'heuristique $HeuristicA$:

$$Best_{Epoch} = min\{(BER_{min} - \frac{1}{2})^2 + Fid_e, pour\ e \in \{1, \ldots, MaxEpoch\}\}, \tag{2.9}$$

où BER_{min} correspond à la valeur minimale du BER obtenue avec les classifieurs externes. Cette heuristique permet de sélectionner la version (l'époque) de l'assainisseur ayant obtenu la plus haute mesure d'équité (mesurée via le BER) et le plus petit dommage, pour chaque paramètre α pris entre $[0, 1]$.

Une fois les différentes époques de l'assainisseur choisies (une version par valeur du paramètre α) sur l'ensemble de validation, nous pouvons mesurer les résultats obtenus sur l'ensemble de test. Nous calculerons, en plus des métriques de protection, les métriques d'utilités et d'équité de groupe Acc_Y, $DemoParity$ et $EqGap$ à partir des mêmes familles de classifieurs externes. Enfin, nous conduirons aussi une analyse détaillée du dommage introduit dans les données par la protection.

2.3 Scénarios d'évaluation

GANSan modifie un ensemble de données d'entrée (incluant l'attribut sensible et la décision) et produit une version assainie (*sans* l'attribut sensible) à partir de laquelle l'inférence de l'attribut sensible est limitée. L'ensemble de données assainies partage donc le même espace que les données originales utilisées en entrée. Dans ce contexte, les performances de GANSan peuvent être évaluées en analysant le front de Pareto caractérisant l'ensemble des possibles compromis entre l'utilité (capturée par la fidélité Fid et la diversité $Diversité$) et l'amélioration de l'équité mesurée par BER et $SAcc$.

Notre approche GANSan peut s'illustrer dans plusieurs scénarios différents. Pour faciliter la compréhension, nous utiliserons les notations suivantes : l'indice tr (respectivement ts) dénote les données de l'ensemble d'entraînement (respectivement de l'ensemble de test). Typiquement, les notations $\{A\}_{tr}$, $\{Y\}_{tr}$, $\{\bar{A}\}_{tr}$ or $\{\bar{Y}\}_{tr}$ seront utilisées pour représenter les valeurs originales des attributs non sensibles de l'ensemble d'entraînement, l'attribut de décision dans ce même ensemble, la version assainie des attributs non sensibles et l'attribut de décision assaini.

Le tableau 2.3 résume la composition de l'ensemble d'entraînement et de l'ensemble de test pour chacun des scénarios d'utilisation de GANSan.

Tableau 2.3 – Scénarios utilisés pour l'évaluation de GANSan. Chaque ensemble est composé des valeurs originales des attributs ou de leurs versions assainies, auquel est rajouté la décision originale ou assainie.

Scenario	**Ensemble d'entraînement**		**Ensemble de Test**	
	A	***Y***	***A***	***Y***
Référence (Baseline)	Original	Original	Original	Original
Scénario 1	Assaini	Assaini	Assaini	Assaini
Scénario 2	Assaini	Original	Assaini	Original
Scénario 3	Assaini	Assaini	Original	Original
Scénario 4	Original	Original	Assaini	Original

Scénario 1 : publication ou assainissement complet de données. Ce cas d'utilisation correspond à un usage typique de l'ensemble de données assainies, qui est la prédiction de l'attribut de décision à l'aide d'un classifieur. Ici l'attribut de décision est, lui aussi, assaini, car nous considérons que cet attribut comporte suffisamment d'information pour inférer l'attribut sensible. Par exemple, on peut

supposer que la décision est obtenue à partir d'un système biaisé envers une population sensible, l'assainisseur est ainsi utilisé pour corriger du mieux possible ce biais. Pour ce scénario, nous quantifierons l'exactitude de prédiction de $\{\bar{Y}\}_{ts}$, ainsi que la discrimination capturée par *DemoParity* (équation 1.5) et *EqGap* (équation 1.8). La limite de ce scénario réside dans le fait que la valeur optimale de la décision exempte de biais à laquelle une personne pourrait se référer est inconnue. Ainsi, dans l'éventualité où la décision serait uniquement obtenue à partir des valeurs de S, le dommage sur Y serait maximal. La décision deviendrait ainsi aléatoire afin de réduire l'influence de S et perdrait donc toute utilité.

Scénario 2 : assainissement partiel des données. L'assainissement partiel est étroitement lié à l'assainissement complet. En effet, dans les deux scénarios, les données d'entraînement et de test sont assainies, à l'exception de la décision qui est maintenue à sa valeur originale durant l'assainissement partiel. Ainsi, pour évaluer ce scénario, les décisions assainies de l'assainissement complet sont remplacées par leurs versions originales. Les ensembles d'entraînement et de test sont donc respectivement composés de $\{\bar{A}_{tr}, Y_{tr}\}$ et $\{\bar{A}_{ts}, Y_{ts}\}$.

L'assainissement partiel est le scénario le plus courant dans la littérature sur l'équité (Edwards et Storkey, 2015 ; Madras, Creager, Pitassi et Zemel, 2018 ; Zemel, Wu, Swersky, Pitassi et al., 2013), car la différence de prédiction de la décision entre un classifieur entraîné sur les données originales $\{A\}_{tr}$ et un autre entraîné sur celles assainies $\{\bar{A}\}_{tr}$ permet de mesurer explicitement la perte d'utilité due à l'assainissement.

Scénario 3 : apprentissage d'un classifieur équitable. Ce scénario a été considéré par Xu et *al.* (Xu, Yuan, Zhang et Wu, 2018) et est motivé par le fait que la modification des données par le processus d'assainissement peut conduire à des résultats illégitimes (modification d'une contravention en crime par exemple dans un dossier pénal) ou encore absurdes (personne âgée de trois ans en maternelle, transformée en personne de trois ans détentrice d'un doctorat). Ainsi, une tierce partie pourrait construire un classifieur non biaisé et utiliser ce classifieur sur des données non perturbées afin de réduire les risques de l'assainissement. Le classifieur serait entraîné sur les données assainies ($\{\bar{A}, \bar{Y}\}_{tr}$) afin d'éviter toutes corrélations avec S, et testé sur les données originales qui sont exemptes de modifications. L'exactitude de prédiction de la décision originale Acc_Y de l'ensemble de test $\{Y\}_{ts}$ et la parité démographique seront utilisées comme mesures de performances.

Scénario 4 : assainissement local. L'assainissement local, comme son nom l'indique, correspond à une

application locale de l'assainisseur par la personne elle-même. L'assainisseur pourrait par exemple être utilisé comme une application déployée sur téléphone mobile, permettant aux personnes de supprimer des corrélations avec l'attribut sensible de leurs profils de données avant de publier celles-ci ou de les partager à différentes entités externes. On pourrait illustrer ce scénario en considérant une entreprise ayant entraîné un classifieur sur des données originales $\{A, Y\}_{tr}$ à des fins de recrutement, ces données originales pouvant être potentiellement biaisées. La personne candidate n'ayant aucun contrôle sur le classifieur, peut toutefois modifier ses données avec l'aide de l'assainisseur et ainsi limiter l'influence de l'attribut sensible de son profil. Le classifieur ne pourrait ainsi pas exploiter des corrélations avec S non désirables. Les performances de l'assainissement sont mesurées dans ce scénario par l'exactitude de prédiction de la décision originale $\{Y\}_{ts}$ ainsi que l'équité quantifiée par $DemoParity$.

L'assainissement local laissant la décision d'assainissement au choix de la personne utilisatrice, ce choix peut conduire à une situation dans laquelle certaines personnes décident de ne pas utiliser l'assainisseur tandis que d'autres l'utiliseraient. Par exemple, les personnes du groupe privilégié peuvent considérer que l'assainissement de leur profil leur ferait perdre l'avantage lié à leur appartenance de groupe, tandis que les groupes protégés peuvent choisir de ne pas révéler leur appartenance pour ne pas subir de discrimination. Ils utiliseraient ainsi l'assainisseur. On pourrait retrouver cet exemple de situation dans le cadre de l'embauche des sapeurs-pompiers. Les personnes de sexe masculin pourraient tirer avantage du fait que ces métiers étaient plus souvent réservés à cette catégorie de personnes, tandis que les personnes de sexe féminin ne souhaiteraient pas révéler cette information pour ne pas subir de biais dans l'évaluation de leur candidature.

2.4 GANSan : résultats

Dans cette section, nous présentons les résultats que nous avons obtenus sur les différents ensembles de données. Dans un premier temps, nous discuterons des résultats généraux sur la protection obtenus dans les deux ensembles, puis des résultats suivant les différents scénarios évoqués en section 2.3. Enfin, nous terminerons par des analyses de fonctionnement de notre approche pour en approfondir la compréhension.

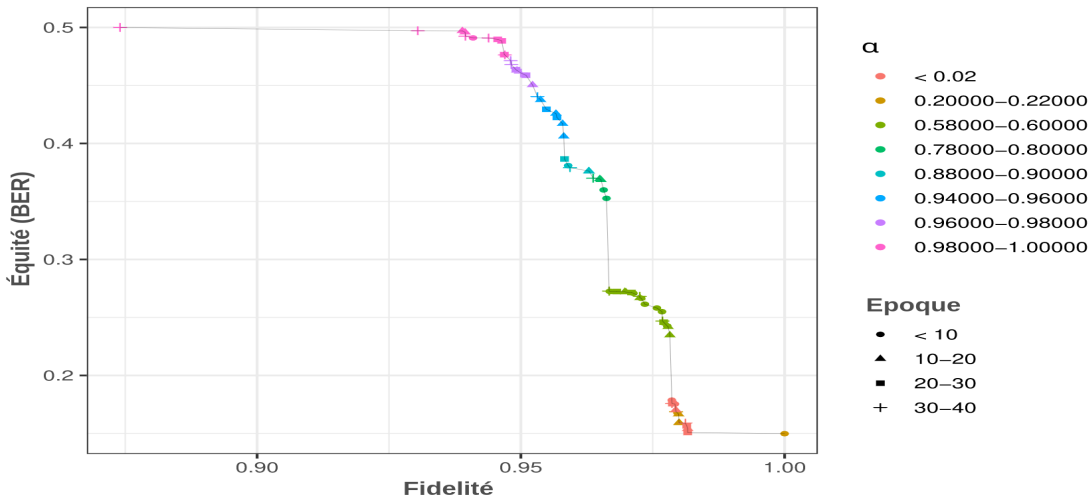


FIGURE 2.3 – Compromis Fidélité-Équité sur l'ensemble de données Adult. Chaque point de la courbe représente le BER minimum possible obtenu à partir des classifieurs externes. L'équité s'améliore avec l'augmentation du coefficient α. Une faible valeur conduit à une petite amélioration de l'équité, tandis qu'une large valeur induit un plus grand dommage dans les données assainies.

2.4.1 Résultats généraux sur Adult Census Income

Nous présentons en figure 2.3 le compromis observé entre l'équité et la fidélité obtenue sur le jeu de données Adult. Nous pouvons observer que l'équité s'améliore avec l'augmentation du coefficient α. L'utilité maximale correspond approximativement à $0,982$ et est atteinte avec $\alpha = 0$, ce qui signifie que l'assainisseur se focalise sur reconstruire les données sans regard à l'équité. Pour $\alpha = 0,2$, la fidélité reste proche de la plus haute valeur atteignable ($Fid_{\alpha=0,2} = 0,98$), mais la protection de l'attribut sensible reste relativement basse ($BER \leq 0,2$). Toutefois, cela correspond à une faible amélioration comparativement aux données originales ($Fid_{orig} = 1$, $BER \leq 0,15$). À l'autre extrême ($\alpha = 1$), les données sont assainies sans considération pour la fidélité par rapport aux données originales. La protection obtenue est par conséquent maximale comme attendu, mais la fidélité est de $Fid_{\alpha=1} \approx 0,88$, soit 10% inférieure à la valeur maximale atteignable. Cependant, un premier compromis intéressant peut être atteint avec $\alpha = 0,96$, qui permet d'atteindre une protection d'approximativement $0,45$ (pour rappel, la protection optimale est de $0,5$) pour un coût pour la fidélité de $2,24\%$ par rapport à la fidélité maximale atteignable ($Fid_{\alpha=0,96} \approx 0,95$).

Concernant la métrique $SAcc$, l'exactitude de prédiction baisse significativement avec l'augmentation

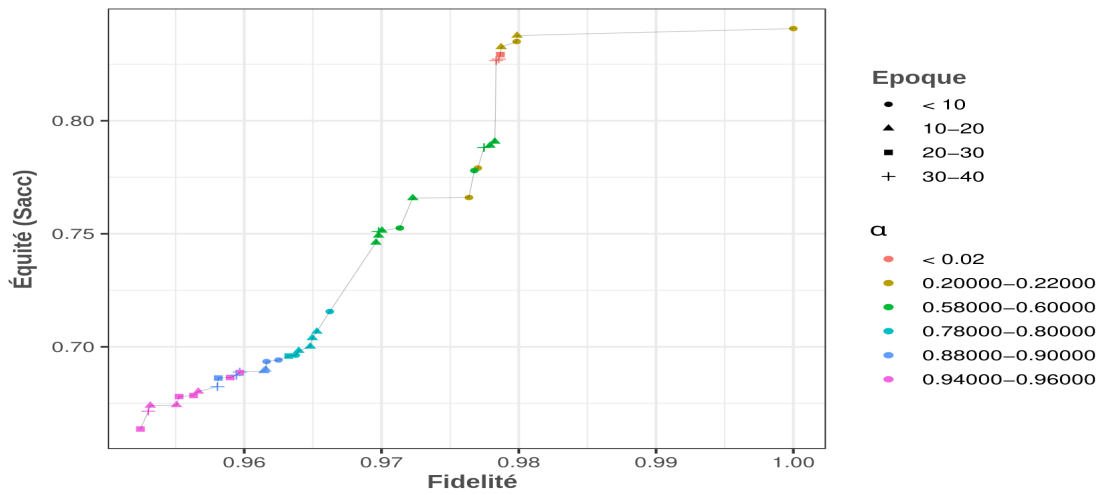


FIGURE 2.4 – Compromis Fidélité-Équité sur l'ensemble de données Adult. Chaque point de la courbe représente le $SAcc$ maximum possible obtenu à partir des classifieurs externes. L'exactitude de prédiction diminue avec l'augmentation de α. Notons la présence d'un dommage minimum introduit par l'assainissement, quelle que soit la valeur de α choisie. Les points de fidélité $Fid = 1$ représentent les valeurs obtenues sur les données originales.

du coefficient α (*cf.* figure 2.4). GANSan réduit les prédictions de S à partir de l'ensemble de données assainies, rendant la métrique proche de la proportion de la classe majoritaire, qui correspond à la proportion du groupe privilégié dans l'ensemble de données. Nous pouvons aussi constater qu'il est impossible d'atteindre la valeur idéale de la protection ($SAcc = 0,6379$), et ce même avec la valeur maximale $\alpha = 1$. De même que le BER, une légère diminution de α ($\alpha = 0,85$) améliore significativement l'assainissement, tout en maintenant une fidélité proche du maximum atteignable.

En figure 2.5, nous résumons l'analyse qualitative au travers de la métrique de diversité *Diversité*. La plus petite perte de diversité observée est de $3,57\%$, qui est atteinte lorsque $\alpha \leq 0,2$. La plus large perte, quant à elle, est de 36%. On peut donc conclure que l'application de GANSan, comme observée au travers de la fidélité avec $\alpha = 0$, introduit de manière systématique des modifications dans l'ensemble de données. La perte de diversité se traduit dans l'ensemble de données par un rapprochement des profils de l'ensemble de données assainies, les rendant de plus en plus similaires avec l'augmentation du coefficient α. Les données sont ainsi réduites à un sous-ensemble restreint de profils typiques. Sur les attributs catégoriques, nous pouvons observer que la proportion des attributs modifiés varie entre 10% et 40% lorsque α est compris dans l'intervalle $[0, 98, 1]$ (*cf.* figure 2.5).

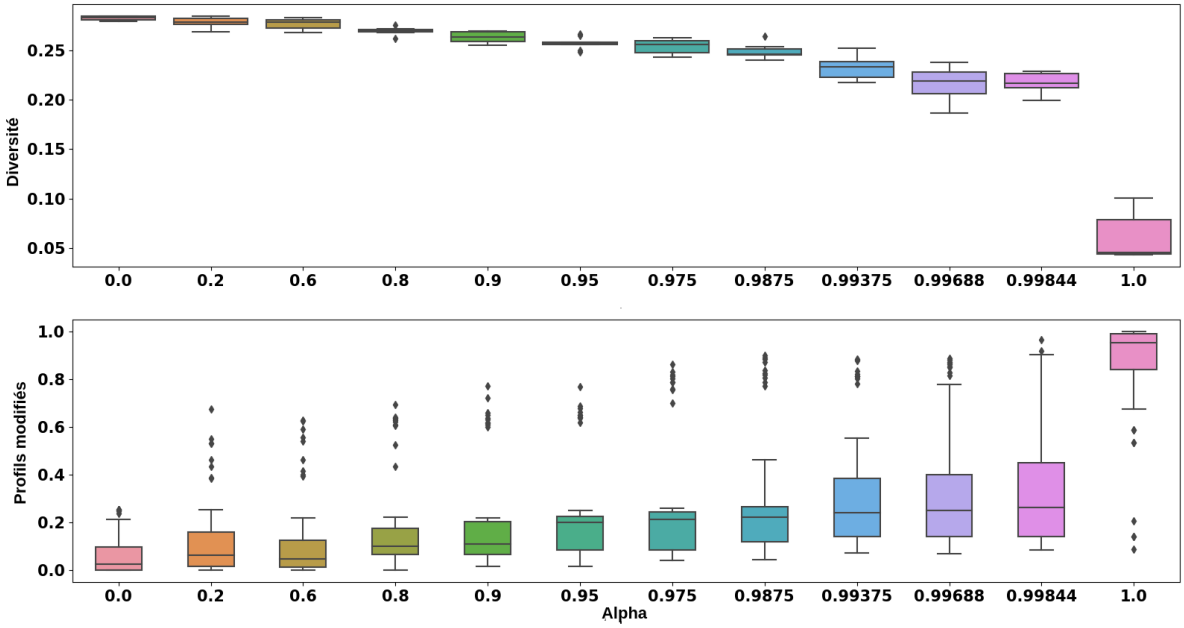


FIGURE 2.5 – Analyse quantitative des ensembles assainis, sélectionnés avec $HeuristicA$. Les métriques sont calculées sur l'ensemble entier (sans subdivision en blocs). Les profils modifiés (*records modified*) correspondent à la proportion de profils ayant des attributs catégoriques modifiés par l'assainissement.

Concernant les attributs numériques, nous avons calculé la métrique de différence relative (*relative difference* or RC en anglais) qui correspond à la différence entre les valeurs originales et assainies normalisées par la moyenne entre les valeurs originales et assainies afin d'éviter des valeurs non existantes, par exemple des divisions par 0.

Définition 18 (Différence relative)**.** *Étant donné deux valeurs numériques originale et assainie, la différence relative mesure l'importance (ou la taille) de la différence absolue entre les valeurs originale et assainie par rapport à une fonction de ces deux valeurs (dans notre cas, la moyenne) est la suivante :*

$$RC = \frac{|originale - assainie|}{f(originale, assainie)} \tag{2.10}$$

$$f(originale, assainie) = \frac{|originale| + |assainie|}{2} \tag{2.11}$$

À l'exception de l'assainissement extrême ($\alpha = 1$), au moins 70% des profils du jeu de données ont un RC inférieure à $0,25$, en grande majorité sur les attributs numériques. Avec $\alpha = 0,9875$, 80% des profils ont un changement relatif inférieur à $0,5$.

2.4.2 Résultats généraux obtenus sur German Credit

De manière similaire à Adult, la protection obtenue sur German Credit a une croissance monotone en fonction de α. La reconstruction maximale ($\alpha = 0$) produit une fidélité d'approximativement $0,96$, tandis que la protection maximale $BER = 0,5$ correspond à une fidélité de $0,81$ et une exactitude de prédiction de $SAcc = 0,76$. Sur la figure 2.6, on observe que la plupart des résultats obtenus stagnent sur un plateau correspondant à une valeur de $SAcc$ de $0,76$, indépendamment de la fidélité et de *alpha*. Nous expliquons ce résultat par le déséquilibre important entre les groupes dans l'ensemble de données.

La protection de l'attribut sensible dans German Credit est très élevée sur l'ensemble de données originelles, avec $BER \approx 0,33$. Néanmoins, on peut observer trois compromis intéressants situés sur les différents coudes du front de Pareto. Nous nommerons ces points A ($BER \approx 0,43, Fid \approx 0,94$), B ($BER \approx 0,45, Fid \approx 0,84$) et C ($BER \approx 0,5, Fid \approx 0,81$), chacun obtenu avec $\alpha = 0,6$ pour le premier, et $\alpha = 0,9968$ pour les deux autres.

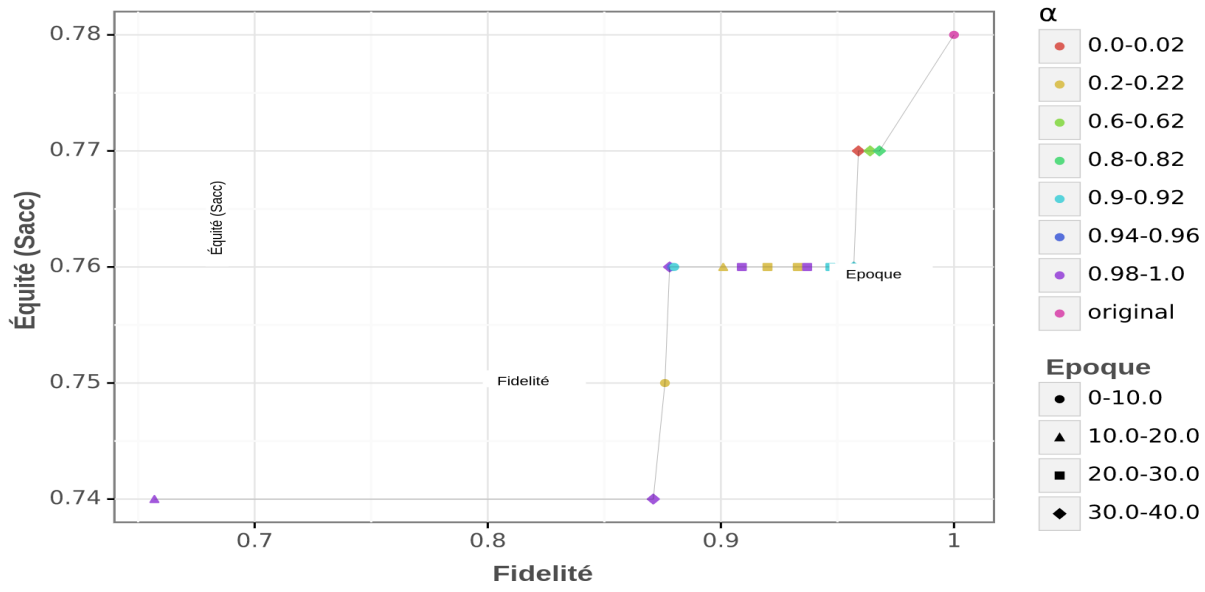


FIGURE 2.6 – Compromis fidelité-équité sur German Credit. Chaque point représente le $SAcc$ maximum possible et obtenu par les classifieurs externes.

Concernant la diversité et le dommage causé par l'assainissement sur les attributs catégoriques de l'ensemble de données German (figure 2.8), nous pouvons constater que la diversité, comme attendue, décroît inversement à *alpha*. Pour $\alpha = 1$ tous les profils sont rendus pratiquement identiques

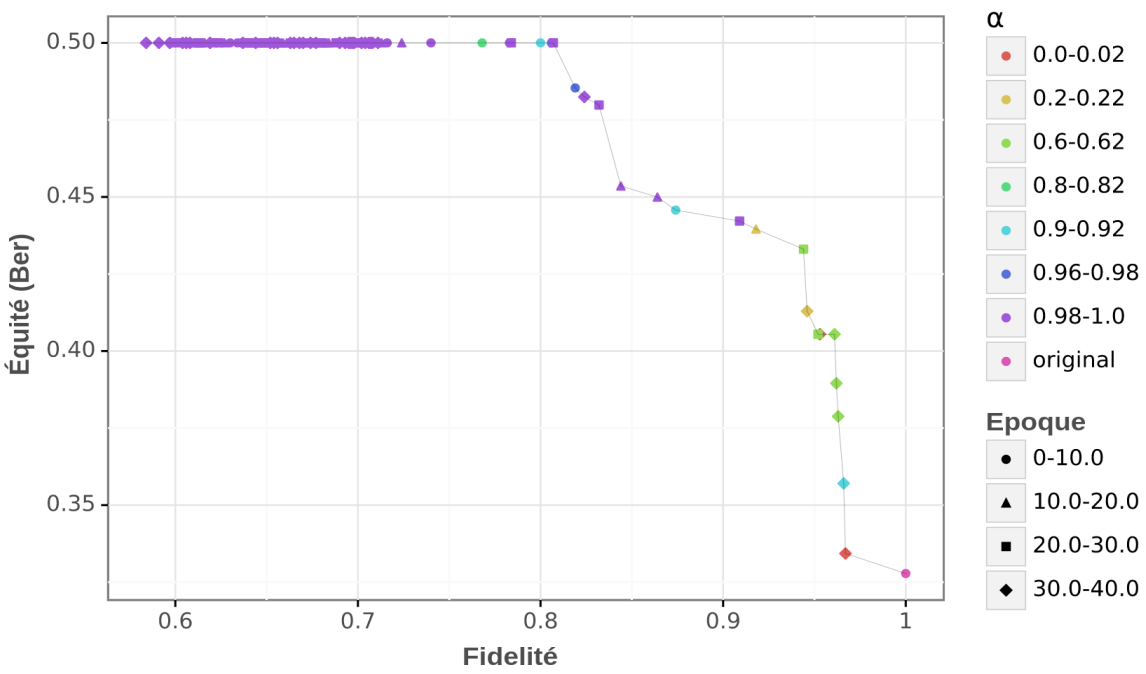


FIGURE 2.7 – Compromis Fidelité-Équité sur German Credit, métrique BER.

(diversité médiane d'approximativement $0,10$, correspondant à $77,77\%$ de perte). Pour les autres valeurs de $\alpha \leq 0,975$, la diversité produit des résultats proches de la valeur originale de $0,51$.

Sur les attributs catégoriques, des tendances similaires à la diversité peuvent être observées : le dommage croît avec α, et les données sont presque totalement détruites lorsque α se rapproche de 1. Pour la plupart des compromis α, le dommage médian est inférieur ou égal à 20%, ce qui signifie que la protection de l'attribut sensible peut être atteinte en modifiant uniquement deux attributs catégoriques du jeu de données. Le changement relatif RC des attributs numériques est inférieur à $0,5$ pour plus de 70% des profils de l'ensemble de données, quelle que soit la valeur de α. Seuls les attributs *Duration in month* et *Credit amount* ont une valeur de dommage plus élevée. Ces dommages peuvent être justifiés par le fait que ces attributs ont des intervalles de valeurs très élevés (33 et 921), particulièrement l'attribut *Credit amount* qui de plus suit une distribution uniforme. Les points de référence considérés A, B et C ont des dommages médians proches de 10% pour A et 20% pour B et C.

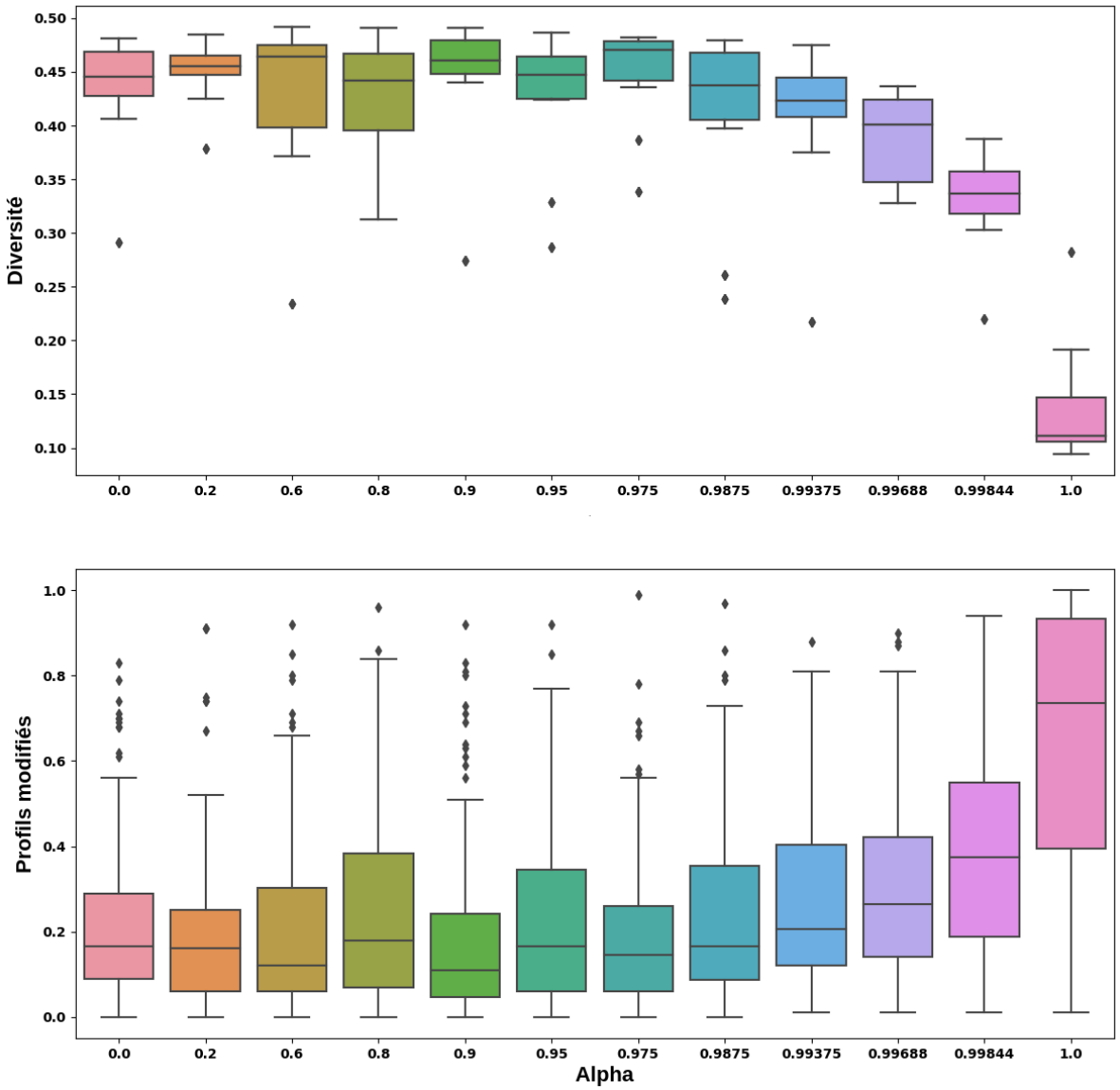


FIGURE 2.8 – Diversité et dommage observé sur les attributs catégoriques pour German Credit.

Pour résumer de manière synthétique nos résultats, notre approche GANSan est en mesure de maintenir la structure de l'ensemble de données à la suite de l'assainissement, rendant les données assainies utiles pour d'autres analyses tout en protégeant l'attribut sensible. Ceci est notamment démontré par les hautes valeurs de fidélité et de protection de S. En conséquence, les résultats obtenus sur les données assainies devraient être proches de ceux calculés sur l'ensemble de données originelles, à l'exception des tâches exploitant des corrélations avec l'attribut sensible S.

Cependant, au niveau individuel (et selon les contextes), certaines modifications peuvent avoir des

conséquences plus importantes que d'autres. Par exemple, en considérant des données concernant des personnes incarcérées, la transformation d'un crime en une contravention pourrait avoir plus d'importance que la modification d'une description physique de la personne. Dans notre dernier travail de recherche *Fair Mapping* (chapitre 4), nous proposons une nouvelle approche pour limiter les conséquences de ces modifications.

Dans la suite de ce chapitre, nous analyserons les résultats obtenus dans chacun des différents scénarios mentionnés précédemment. Pour cela, nous avons fixé la valeur de α à $0,9875$ et $0,9938$, ce qui correspond à un niveau de protection proche de l'optimal ($0,4897$ et $0,4892$ respectivement), tout en maintenant un niveau acceptable de dommage dans les données ($Fid_{\alpha=0,9875} \approx 0,9464$ and $Fid_{\alpha=0,9938} \approx 0,9425$). Nous limiterons la présentation de nos résultats à ceux obtenus sur l'ensemble de données Adult, mais tous les résultats détaillés sont présentés dans notre article (Aïvodji, Bidet, Gambs, Ngueveu et al., 2021).

2.4.3 Scénario 1 : assainissement complet

L'objectif de ce scénario est de mesurer l'utilité préservée dans l'ensemble de données après un assainissement de toutes les données. L'utilité ici sera mesurée par l'exactitude de prédiction de la décision Y.

Nous pouvons observer que l'assainissement préserve bel et bien l'exactitude de prédiction de la décision, qui plus est, l'assainissement améliore la capacité de prédiction de la décision pour tous les classifieurs, comme présentée en figure 2.9, *Scenario S1* (Acc_Y égal $0,86$, $0,84$ et $0,78$, respectivement pour les classifieurs GB, MLP et SVM). L'amélioration de la prédiction s'explique par le fait que l'assainissement produit par GANSan décorrèle la décision de l'attribut sensible, et la transforme pour la rendre plus cohérente avec les autres attributs descriptifs du profil. L'impact du biais est ainsi limité, les profils similaires (indépendamment du groupe d'appartenance) étant associés à des décisions proches. En effet, en assainissant la décision, on observe dans l'ensemble de données assainies une distribution de la décision similaire à celle du jeu de données original, à l'exception de quelques profils dont la décision a été inversée ($7,56\% \pm 1,23\%$ de profils ont leur décision changée dans l'ensemble de données, dont $11,44\% \pm 2,74\%$ dans le groupe protégé pour $\alpha = 0,9875$). La variation de la décision peut s'expliquer par la similarité entre les profils du groupe protégé et ceux du groupe privilégié, qui toutefois ont des décisions opposées. En effet, en observant les taux de décisions positives dans les deux groupes avant et après assainissement (tableau 2.4),

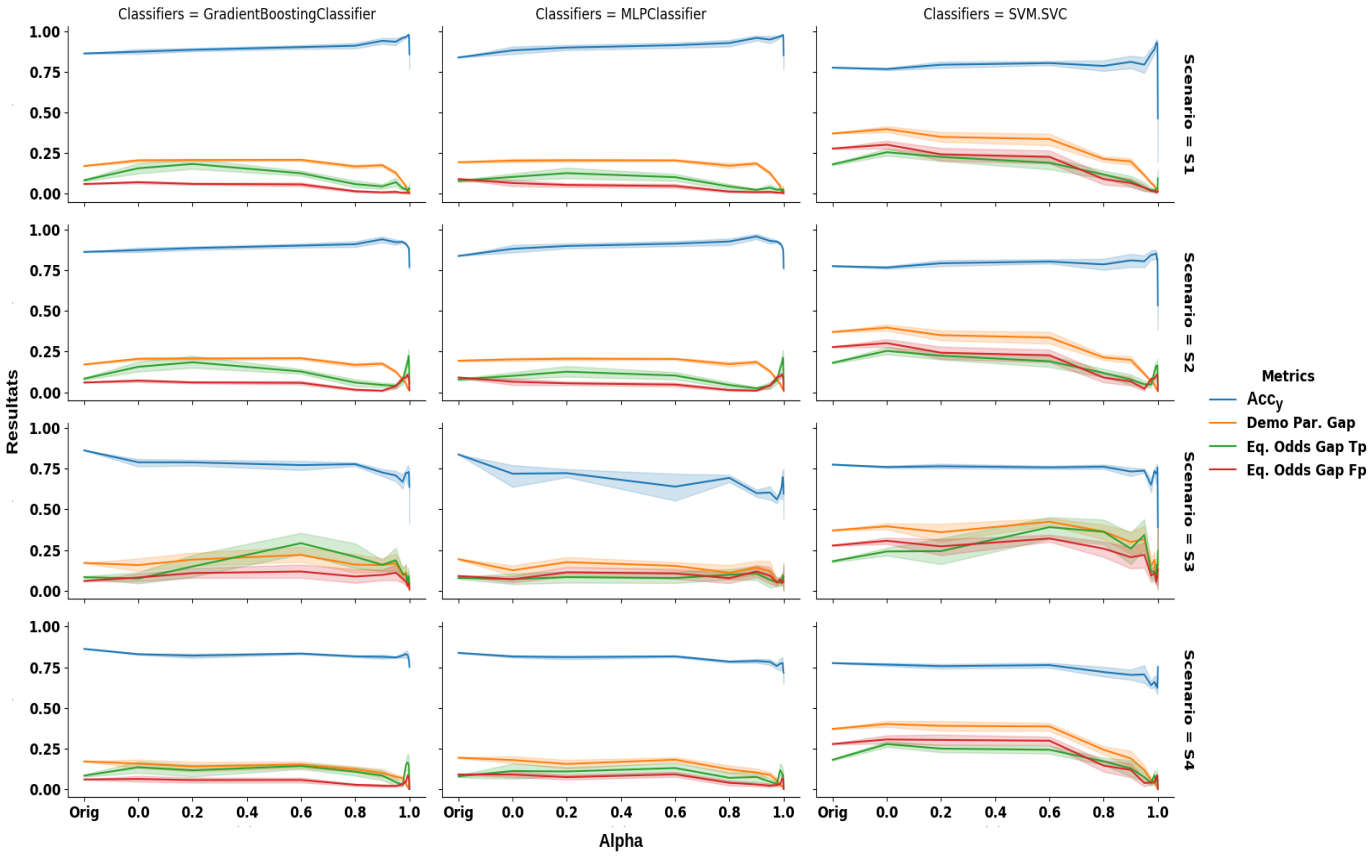


FIGURE 2.9 – Exactitude de prédiction (bleu), parité démographique (orange) et égalité des chances (vrais positifs en vert et faux positifs en rouge) calculées sur les scénarios 1, 2, 3 et 4 (de haut en bas) avec les classifieurs GB, MLP et SVM (de gauche à droite) sur l'ensemble de données Adult. Plus la valeur de α est élevée, meilleure est l'équité. L'utilisation des données assainies $\bar{A}$ (tel que dans le scénario S1 et S2) améliore l'exactitude de précision, tandis que l'utilisation de la combinaison entre les données assainies $\bar{A}$ et originales A la dégrade.

on constate que ce taux augmente dans le groupe protégé pour se rapprocher de celui du groupe privilégié privilégié, tandis qu'il ne varie pas dans la distribution privilégiée. L'amélioration de la

TABLEAU 2.4 – Taux de décision positives prédites avec le classifieur GB.

	Protégée	Privilégiée
Originale	$\approx 0,08$	$\approx 0,25$
Assainie	$\approx 0,2$	$\approx 0,25$

prédiction de la décision peut aussi s'expliquer par la légère perte en diversité. En effet, comme nous l'avons mentionné précédemment, une perte en diversité indique une plus grande similarité entre les profils, et par conséquent, une prédiction plus aisée de la décision.

Concernant la discrimination mesurée par $DemoParity$, $EqGap_0$ et $EqGap_1$, nous pouvons observer que toutes ces métriques sont des fonctions décroissantes du compromis α. Avec $\alpha \geqslant 0,6$, les corrélations avec S sont significativement réduites. Il en va de même pour les écarts mesurés par la parité démographique et l'égalité des chances.

Pour $\alpha = 0,9875$, $BER \geq 0,48$, $Acc_Y = 0,965$, $DemoParity = 0,0453$, $EqGap_1 = 0,0286$ and $EqGap_0 = 0,0062$ for GB tandis que sur les données originales, nous obtenons respectivement, $DemoParity = 0,16$, $EqGap_1 = 0,083$ $EqGap_0 = 0,060$. Les performances de GANSan sont encore améliorées avec $\alpha = 0,9938$: $BER \geq 0,48$, $Acc_Y = 0,973$, $DemoParity = 0,0185$, $EqGap_1 = 0,0161$ and $EqGap_0 = 0,0045$.

Dans ce scénario, FairGan (Xu, Yuan, Zhang et Wu, 2018) arrive à un BER de $0,3862 \pm 0036$, une exactitude de prédiction de $0,8247 \pm 0,0115$ et une parité démographique correspondant à $0,0354 \pm 0,0206$ alors que FairGan+ (Xu, Yuan, Zhang et Wu, 2019) offre les performances suivantes sur la protection de S, BER de $0,3867 \pm 0049$, une exactitude de prédiction de $0,817 \pm 0,003$ et une parité démographique de $0,014 \pm 0,0065$.

2.4.4 Scénario 2 : assainissement partiel

Dans ce scénario, nous mesurons l'équité et l'utilité préservée par l'assainissement de toutes les caractéristiques des profils, à l'exception de la décision.

Puisque l'assainissement supprime les corrélations avec la décision, et introduit des perturbations dans les données, on pourrait s'attendre à ce que l'exactitude de précision soit inférieure à sa valeur calculée sur les données originales. Contrairement à nos attentes, nous pouvons observer que l'exactitude de la précision s'améliore avec l'augmentation des valeurs de α. La parité démographique s'améliore elle aussi, tandis que l'égalité des chances reste pratiquement constante ($EqGap_1$, ligne verte sur la figure 2.9).

Dans le tableau 2.5, nous faisons une comparaison entre les résultats obtenus par notre approche et ceux de l'état de l'art. Les résultats présentés sont ceux obtenus avec le classifieur ayant la plus haute exactitude de prédiction (MLP) et la plus basse (SVM).

TABLEAU 2.5 – Comparaison des approches sur la base de l'exactitude de prédiction et la parité démographique sur le jeu de données Adult.

Method	Acc_Y	$DemoParity$	$EqGap_1$	$EqGap_0$
LFR (ZEMEL, WU, SWERSKY, PITASSI et al., 2013)	$0,78$	$\approx 0,02$	–	–
ALFR (EDWARDS et STORKEY, 2015)	$0,825$	$\approx 0,02$	–	–
MUBAL (ZHANG, LEMOINE et MITCHELL, 2018)	$0,845$	$0,1$	$0,0108$	$0,0053$
LATR (MADRAS, CREAGER, PITASSI et ZEMEL, 2018)	$0,84$	$0,1$	-	$0,029$
FairGan (XU, YUAN, ZHANG et WU, 2018)	$0,8256 \pm 0,0021$	$0,0901 \pm 0,0220$	$0,1473 \pm 0,0608$	$0,0361 \pm 0,0145$
$FairGan+_{DP}$ (XU, YUAN, ZHANG et WU, 2019)	$0,8178 \pm 0,0035$	$0,0141 \pm 0,0065$	-	-
$FairGan+_{EO}$ (XU, YUAN, ZHANG et WU, 2019)	$0,8218 \pm 0,0062$	-	$0,0312 \pm 0,0316$	$0,0245 \pm 0,0124$
GANSan *(S2) - MLP,* $\alpha = 0,9875$	$0,9143 \pm 0,0136$	$0,0508 \pm 0,0253$	$0,1249 \pm 0,0668$	$0,0975 \pm 0,0313$
GANSan *(S2) - SVM,* $\alpha = 0,9875$	$0,8489 \pm 0,0476$	$0,0480 \pm 0,0258$	$0,1473 \pm 0,0664$	$0,0830 \pm 0,0293$
GANSan *(S2) - MLP,* $\alpha = 0,9938$	$0,9003 \pm 0,0111$	$0,0283 \pm 0,0154$	$0,1769 \pm 0,0402$	$0,1086 \pm 0,0289$
GANSan *(S2) - SVM,* $\alpha = 0,9938$	$0,8536 \pm 0,0433$	$0,0214 \pm 0,0165$	$0,1612 \pm 0,0497$	$0,1019 \pm 0,0310$

À partir du tableau, nous pouvons observer que GANSan a de meilleures performances que les autres approches, notamment sur l'exactitude de prédiction, mais le plus petit niveau de parité démographique est obtenu avec FairGan+ (XU, YUAN, ZHANG et WU, 2019) ($DemoParity = 0,014$). Ceci s'explique notamment par le fait que cette approche a été spécialement conçue pour améliorer cette métrique. Notre approche, de même que FairGan (XU, YUAN, ZHANG et WU, 2018) a des performances limitées lorsque la décision originale est utilisée (métrique $EqGap$). Ces faibles performances peuvent être expliquées par le fait que les corrélations entre S et les décisions originelles ont été supprimées par l'assainissement du jeu de données. Ainsi, les prédictions obtenues à partir des données assainies ne sont pas alignées sur les prédictions originales. Nous pouvons aussi observer que notre approche permet d'obtenir l'un des meilleurs résultats sur la parité démographique. FairGan+ (ZHANG, LEMOINE et MITCHELL, 2018) et MUBAL (ZHANG, LEMOINE et MITCHELL, 2018) produisent les meilleurs résultats mesurés par l'égalité des chances, ce qui est attendu, puisque ces approches ont été spécifiquement conçues pour améliorer cette métrique.

Pour conclure bien que ces mesures d'équité ne fassent pas partie des objectifs optimisés explicitement, nous pouvons observer que notre approche les améliore considérablement dans ce scénario dans lequel les décisions peuvent être corrélées à l'attribut sensible.

2.4.5 Scénario 3 : apprentissage d'un classifieur équitable

L'assainisseur peut être utilisé pour réduire la discrimination lors d'une tâche de classification utilisant les données originales, à partir du moment où le classifieur est entraîné sur les données assainies. La troisième ligne de la figure 2.9 présente les résultats obtenus dans le cadre de ce scénario, où nous pouvons observer que la discrimination (lors de la tâche faisant appel aux données originales) est réduite. Cependant, on peut aussi observer que la courbe représentant Acc_Y en fonction de α affiche la plus grande pente négative de tous les scénarios. Plus précisément, le meilleur classifieur à prédire la décision (sur les données originales) affiche une perte d'approximativement 16%, ce qui peut se justifier par la différence de corrélations entre A et Y, et entre $\bar{A}$ et $\bar{Y}$. En effet, le classifieur est entraîné sur les données assainies ($\bar{A}$ et $\bar{Y}$), et de ce fait la frontière de décision apprise est moins appropriée pour les données originales ($\bar{A}$ et $\bar{Y}$), qui utilisent les corrélations avec S.

Ce scénario a aussi été utilisé par FairGan (Xu, Yuan, Zhang et Wu, 2018). Leurs performances sont notamment de $Acc_Y = 0,82$ et $DemoParity = 0,0461 \pm 0,0424$ tandis que les nôtres obtenues avec le classifieur GB sont de $Acc_Y = 0,724 \pm 0,038$ et $DemoParity = 0,111 \pm 0,059$ pour $\alpha = 0,9875$ et de $Acc_Y = 0,725 \pm 0,107$ et $DemoParity = 0,0598 \pm 0,0422$ pour $\alpha = 0,9938$.

2.4.6 Scénario 4 : assainissement local

Notre première observation dans ce scénario est la diminution de la discrimination avec l'augmentation du coefficient α. De même que pour les autres scénarios, plus les corrélations avec l'attribut sensible sont enlevées, plus importante est l'amélioration en équité (mesurée par $DemoParity$, $EqGap_1$ et $EqGap_0$), et plus petite est l'exactitude de prédiction de l'attribut de décision original. Par exemple, avec GB, nous obtenons $Acc_Y = 0,83 \pm 0,039$ et $DemoParity = 0,035 \pm 0,022$ lorsque $\alpha = 0,9875$ et $Acc_Y = 0,8240 \pm 0,0352$ et $DemoParity = 0,0114 \pm 0,0061$ lorsque $\alpha = 0,9938$ (les valeurs sur l'ensemble de données originales sont de $Acc_Y = 0,86$ et $DemoParity = 0,16$).

Nous avons aussi mesuré les performances de GANSan dans ce scénario respectivement par rapport aux décisions assainies $\bar{Y}$ au lieu des décisions originales comme précédemment. En d'autres termes, le classifieur est entraîné sur les données originales et nous avons mesuré par exemple l'exactitude de prédiction en utilisant les décisions assainies. Dans ce contexte, les performances de GANSan sont significativement améliorées. En effet, l'exactitude de prédiction avec GB atteint la valeur de $Acc_{\bar{Y}} = 0,8703 \pm 0,0589$ avec $\alpha = 0,9938$, tandis que l'égalité des chances varie de $EqGap_1 = 0,1646 \pm 0,0927$

et $EqGap_0 = 0,0853 \pm 0,0319$ obtenues avec la décision originale, à $EqGap_1 = 0,0243 \pm 0,0201$ and $EqGap_0 = 0,0084 \pm 0,0075$ pour les décisions assainies. Comme mentionné dans le scénario *S2*, ceci signifie que les corrélations avec les décisions originales ne sont pas préservées par le processus d'assainissement. Notons que la métrique *DemoParity* reste inchangée puisque celle-ci ne nécessite que les prédictions réalisées par le classifieur.

Ces résultats montrent la possibilité pour GANSan d'être utilisé localement, permettant ainsi aux personnes (sans l'intervention d'une tierce partie) de pouvoir partager leur information. L'attribut sensible pour lequel GANSan a été entraîné à protéger est masqué. De même, la légère baisse en exactitude de prédiction ($3,68\%$ avec GB et 8% pour MLP) rend l'approche GANSan encore plus intéressante dans la mesure où elle permet l'amélioration des métriques de protection et d'équité sans nécessiter un re-entraînement des modèles. Ainsi, pour des structures et des réseaux de neurones beaucoup plus complexes et nécessitant beaucoup de ressources matérielles et temporelles, GANSan offre un compromis intéressant entre équité, utilité et complexité d'implémentation.

2.4.7 Amélioration des prédictions par l'assainissement

Dans le scénario 1 (assainissement complet), nos résultats ont montré que l'assainissement peut améliorer la capacité d'un classifieur à prédire la décision. En plus des possibles hypothèses émises (perte en diversité, similarité augmentée entre profils), ces observations indiquent que l'assainissement transforme les données de sorte que les caractéristiques descriptives des profils de personnes sont «alignées» sur la distribution des attributs, ainsi que sur les distributions conditionnelles obtenues par la combinaison des valeurs des attributs.

Pour illustrer nos propos, considérons un ensemble de données dans lequel les profils sont composés de l'attribut sensible binaire *genre* (ayant pour valeurs S_0 et S_1), un attribut *occupation* et d'autres attributs nommés X que nous supposerons indépendant de l'attribut S. De plus, considérons que 80% des profils dans le groupe S_0 ont pour occupation *comptable*, tandis que les autres profils de l'ensemble ont la valeur $professeur$. Dans cet exemple, nous pouvons constater qu'un classifieur entraîné à prédire l'occupation serait en mesure de le réaliser facilement si l'attribut *genre* est inclus en intrant (étant donné que cet attribut est fortement corrélé à l'appartenance de groupe). Le processus d'assainissement appliqué à cet ensemble de données modifiera la distribution conditionnelle de l'attribut *occupation* (puisque X sont supposés indépendants de S) en prenant compte des corrélations entre cet attribut et S, ainsi que de l'alignement qui doit être fait entre les distribu-

tions conditionnelles de cet attribut entre les groupes (protégé et privilégié). En d'autres termes, la protection de l'attribut sensible demanderait à ce que $Pr(S_0|occupation) = Pr(S_1|occupation)$, tandis que l'alignement voudrait $Pr(X = x|occupation = o, S_0) = Pr(X = x|occupation = o, S_1)$, avec $o \in \{comptable, professeur\}$, et x une valeur de l'attribut X. Si X correspond à l'attribut de décision dans ce jeu de données, on constate ainsi que l'alignement modifierait la valeur de l'attribut *occupation* de sorte les distributions conditionnelles $Pr(X = x|occupation = o, S_0)$ et $Pr(X = x|occupation = o, S_1)$ soient égales et par conséquent, que les valeurs du profil soient alignées pour conduire aux mêmes décisions. Dans le cas où X et S ne seraient pas indépendants, l'assainissement modifierait les deux attributs X et *occupation*, tout en essayant d'aligner les distributions $Pr(X = x|occupation = o, S_0)$ et $Pr(X = x|occupation = o, S_1)$. Un processus similaire, mais plus complexe pourrait être réalisé par GANSan sur les données de plus hautes dimensions.

Cette observation permettrait aussi d'inférer le fait que l'assainissement ne conduirait pas nécessairement à une diminution de l'utilité de l'ensemble de données assaini (notamment la prédiction de la décision, bien que le dommage sur cet attribut soit important), et ce, même si l'attribut de décision est fortement corrélé avec l'attribut sensible. Toutefois, la procédure d'alignement des valeurs de l'assainissement ne tient pas compte de la signification sémantique des combinaisons de valeurs d'attributs, mais se limite uniquement aux distributions.

Pour approfondir notre analyse, nous avons créé un ensemble de données synthétiques, avec une proportion égale de profils dans les différents groupes. L'attribut sensible ici est le *genre* (*gender*, valeurs 0 et 1). Nous considérons trois attributs numériques, $note_1$, $note_2$ et $note_3$ et un attribut de décision *noteDec*. Chaque attribut numérique provient d'une distribution gaussienne centrée sur $mean = 15$ et d'écart-type $std = 1$. La décision pour chaque ligne est générée en prenant la moyenne des trois attributs numériques (M_3) et elle est biaisée envers le groupe 0. Le biais consiste en l'application de seuils différents : pour le groupe $gender = 1$, les décisions positives sont obtenues avec l'équation $M_3 \geq mean$; pour le groupe $gender = 0$, les décisions sont considérées positives si $M_3 \geq mean + 0,7$. La décision est ainsi corrélée à l'attribut sensible tandis que les autres en sont indépendants. Notre hypothèse sur l'alignement stipule que l'attribut de décision (étant le seul corrélé à l'attribut sensible, et ainsi à modifier) sera modifié de sorte que $P(noteDec = 1|note_1, note_2, note_3, gender = 0) = P(noteDec = 1|note_1, note_2, note_3, gender = 1)$, et $P(noteDec = 0|note_1, note_2, note_3, gender = 0) = P(noteDec = 0|note_1, note_2, note_3, gender = 1)$. L'alignement implique ainsi que l'assainissement n'est pas un processus aléatoire qui introduit uniquement des perturbations limitées. La protec-

tion de l'attribut sensible impliquerait $P(gender = 0|noteDec = 0) = P(gender = 1|noteDec = 0)$ et $P(gender = 0|noteDec = 1) = P(gender = 1|noteDec = 1)$, ce qui se traduit par une même valeur de seuil pour la décision dans les deux groupes.

En figure 2.10 l'ensemble de données assaini a le même niveau de protection (BER) que l'ensemble de données original. Contrairement à nos attentes, le processus d'assainissement n'a pas modifié l'attribut de décision, mais plutôt l'attribut $note_1$ de sorte que $noteDec$ soit le résultat d'une fonction unique appliquée à l'ensemble du jeu de données (figure 2.11), les distributions conditionnelles des deux groupes affichent aussi une certaine superposition.

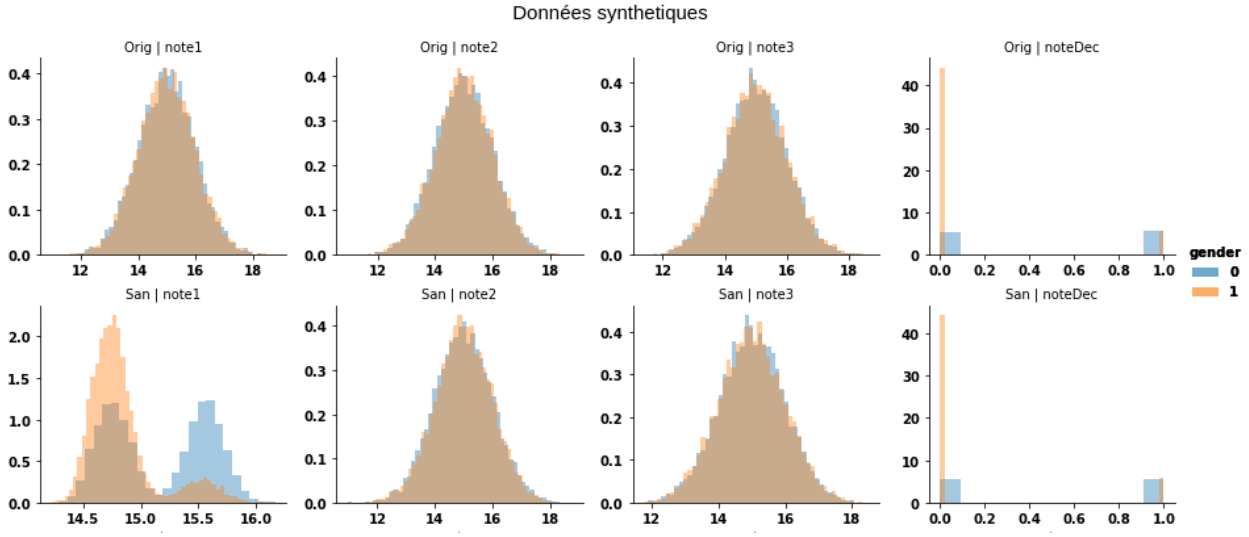


FIGURE 2.10 – Distribution des attributs de l'ensemble de données. De haut en bas : distributions originales puis distributions assainies. De gauche à droite : $note_1$, $note_2$, $note_3$ et $noteDec$. Les données originales montrent la similarité entre les distributions des attributs, mais des décisions différentes. L'assainissement commence par aligner la distribution $note_1$ de sorte que cette dernière suit la distribution de la décision, l'attribut sensible n'a pas encore été protégé.

Des observations similaires sur l'*alignement* peuvent être faite sur un ensemble de données similaire, mais dans lequel la discrimination a été empirée (le seuil de décision positive pour le groupe $gender = 0$ a été augmenté tandis que nous l'avons réduit pour le groupe $gender = 1$). L'attribut $note_3$ a été rendu identique pour les deux groupes par l'assainissement (figure 2.12), bien que la prévention d'inférence de l'attribut sensible et la qualité de reconstruction ne sont pas encore près de l'optimal.

En augmentant la qualité de protection de l'attribut sensible sur ce jeu de données, l'assainissement

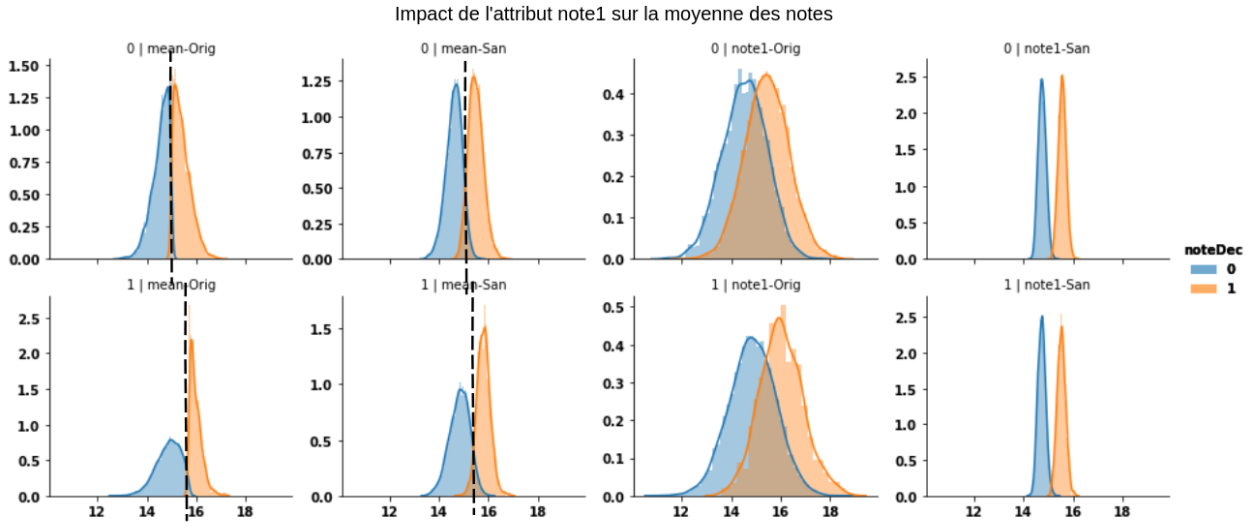


FIGURE 2.11 – Frontière de décision basée sur la moyenne des trois attributs $note_x$ sur les données originales (deuxième plus à gauche) et assainies (gauche) pour le groupe $gender = 0$ (haut) et $dender = 1$ (bas). L'assainissement modifie les frontières de décision pour les chacun des groupes, les rendant presque identiques. *Orig* fait référence aux distributions originales tandis que *San* identifie leurs versions assainies.

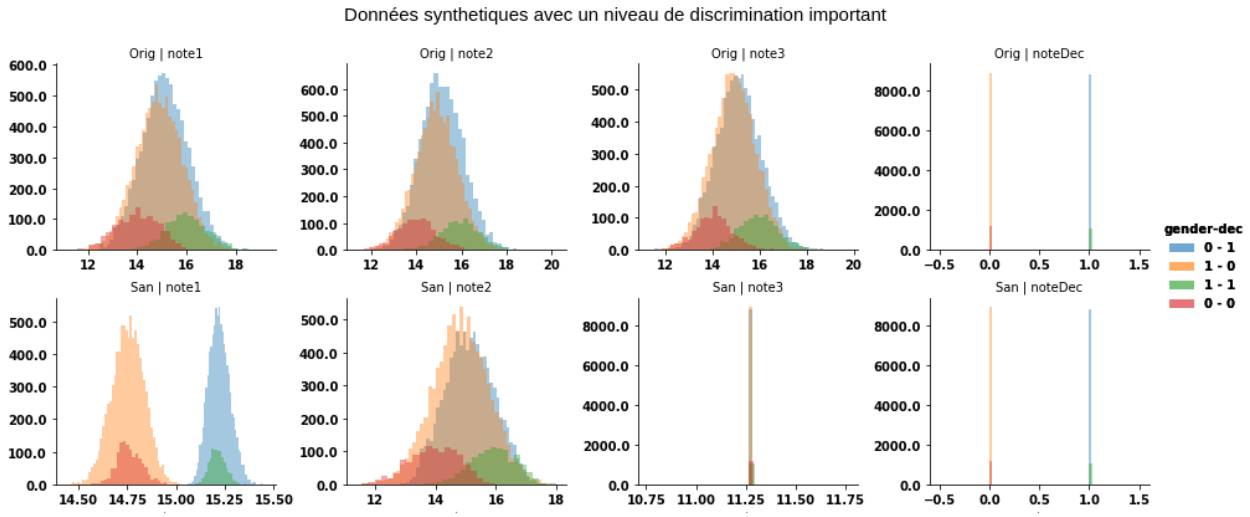


FIGURE 2.12 – Distribution des attributs sur l'ensemble de données synthétiques, mais avec une discrimination plus importante. L'assainissement commence par l'alignement de quelques distributions ($note_1$) de sorte qu'elles correspondent au critère de décision *noteDec*.

triple la protection mesurée par le BER, en modifiant tous les attributs pour protéger l'attribut sensible. La similarité entre les distributions est maintenue, tandis que la déviation est réduite (Figure 2.13).

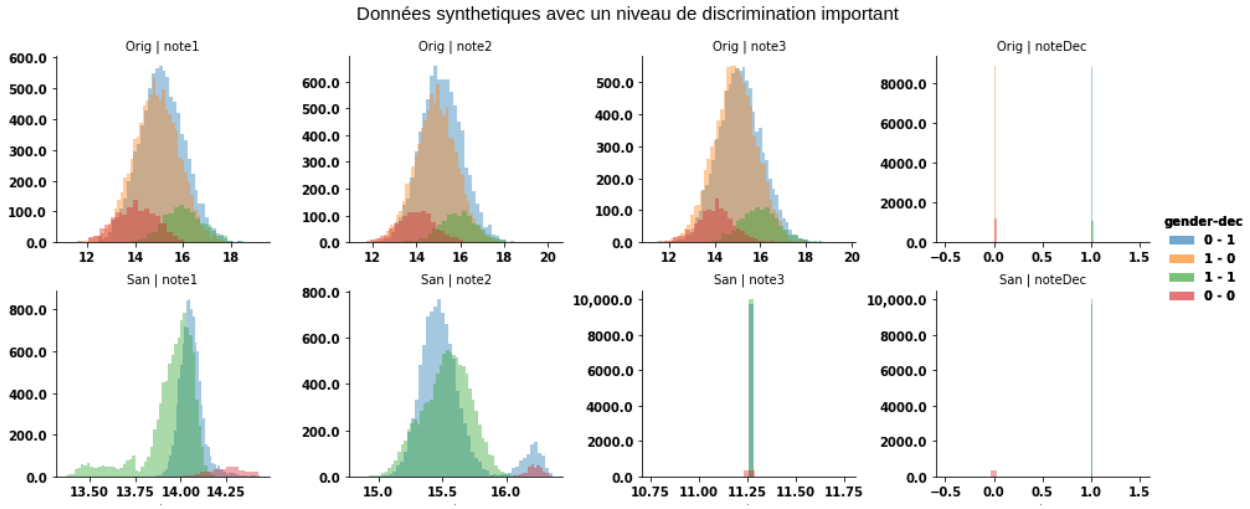


FIGURE 2.13 – Distribution des attributs sur l'ensemble de données synthétiques, avec une protection accrue de l'attribut sensible (le BER passe de $0,1$ à $0,32$).

L'alignement permet ainsi d'expliquer l'amélioration de la qualité de prédiction de certains attributs, ainsi que la discordance entre les décisions assainies et leurs versions originales.

2.5 Temps d'exécution de GANSan

Nous avons comparé le temps d'exécution de notre approche avec d'autres de l'état de l'art en utilisant des cadriciels disponibles. GANSan et FairGan sont implémentées avec le cadriciel Pytorch (PASZKE, GROSS, MASSA, LERER et al., 2019), l'approche *Disparate Impact Remover* (DIRM) (FELDMAN, FRIEDLER, MOELLER, SCHEIDEGGER et al., 2015) et Learning Fair Representation (LFR) (ZEMEL, WU, SWERSKY, PITASSI et al., 2013) (avec les hyperparamètres $K = 50, A_x = 0,01, A_y = 1, A_z = 50$) ont été récupérées du cadriciel AIF360 (BELLAMY, DEY, HIND, HOFFMAN et al., 2018). Les temps d'exécution ont été mesurés sur un ordinateur de capacité (*Intel Core i7-8750H CPU @ 2,20 GHz* avec 30 Gi), en utilisant l'ensemble de données *Adult Census*. Pour accélérer les calculs, certaines de nos expériences ont été réalisées sur la plateforme *Calcul Canada* [4] qui offre de meilleures

4. `https://cc.sillmedia.com/fr/`

ressources.

DIRM est l'approche la plus rapide ne nécessitant que $9,8s$ alors que LFR s'exécute en $2563,33s$. Concernant FairGan, nous avons évalué le temps d'exécution d'une époque à l'étape d'auto-encodage $(5,3s)$, d'apprentissage de la distribution $(2,75s)$ et d'amélioration de l'équité $(5,19s)$. Ainsi, en utilisant les paramètres disponibles dans l'article original, FairGan nécessite $16960s$. GANSan nécessite $460,37s$ par époque, donc $18414,8s$ pour les 40 époques de nos expériences. Le surcoût de GANSan est principalement dû à l'optimisation sous forme de vecteurs, ainsi que la largeur des réseaux utilisés. Plus précisément, les couches utilisées par le discriminateur sont respectivement des matrices de taille $(data\_input\_shape, data\_input\_shape * 16)$ pour la première couche, et $(data\_input\_shape * \frac{16}{2^{i-2}}, data\_input\_shape * \frac{16}{2^{i-1}})$ pour les couches i subséquentes alors que la couche de sortie est de $(data\_input\_shape * 2, output\_size)$. Cette structure du discriminateur est celle qui fournit empiriquement de meilleurs résultats tout en nécessitant un temps d'exécution raisonnable. En utilisant la même structure que FairGan, notre approche s'exécute en $158,486$ seconds par époque ($6339s$ pour 40 époques).

2.6 GANSan : conclusion

En résumé, GANSan est une approche d'assainissement des données inspirée par les RAG, et qui améliore l'équité par la prévention d'inférence de l'attribut sensible à partir d'autres informations présentes dans le jeu de données. Nos résultats montrent que GANSan est efficace dans la protection de l'attribut sensible, en plus de limiter la perte en utilité due à l'assainissement, notamment en ce qui concerne l'exactitude de prédiction d'une décision, et le dommage mesuré sur les attributs numériques et catégoriques. GANSan présente aussi l'avantage d'offrir la possibilité d'assainissement local, en modifiant le moins possible les attributs tout en préservant l'espace des données originales (ce qui maintient la compréhensibilité des données). La protection offerte par GANSan repose sur l'incapacité du discriminateur à inférer l'attribut sensible, qui est aussi mesuré par trois différents types de classifieurs externes. Toutefois, un classifieur plus performant ou mieux configuré pourrait arriver à inférer S avec de meilleures performances. Néanmoins, cette limitation est inhérente à la plupart des approches de prétraitement des données et non uniquement la nôtre.

Comme nous l'avons mentionné dans le chapitre 1, la protection de l'attribut sensible trouve un écho dans la protection de la vie privée. Dans le chapitre 3, nous montrons comment GANSan a été adapté pour la protection des données de santé. Malgré les performances de GANSan, plusieurs

limitations demeurent, dont notamment l'aspect binaire de l'attribut sensible, la perte de signification sémantique des données assainies ou encore l'introduction de modifications illégitimes qui rendent l'approche difficilement utilisable dans des contextes critiques. Dans le chapitre 4, nous tenterons de résoudre certains de ces limitations avec l'approche FairMapping.

CHAPITRE III

DYSAN: ASSAINISSEMENT DYNAMIQUE DES DONNÉES DE CAPTEURS PAR DES RÉSEAUX ANTAGONISTES

Notre approche GANSan qui sert à protéger les informations personnelles a été aussi adaptée dans le contexte de la protection des informations privées provenant du domaine de la santé. Pour ce cas d'application, nous avons considéré une utilisatrice (Alice), qui utiliserait des capteurs connectés ou son smartphone pour mesurer divers paramètres physiologiques et notamment surveiller son activité physique. L'application effectue la reconnaissance de l'activité ainsi que le suivi de l'activité dans le nuage. Autrement dit, même si le fournisseur de services le déclare explicitement dans les conditions générales du service, cette utilisatrice n'a aucune garantie formelle que ses données ne seront pas exploitées pour déduire d'autres informations sur son sujet (par exemple, à des fins de ciblage ou de marketing). Un autre scénario possible est lié à la nouvelle tendance des compagnies d'assurance qui proposent des réductions aux clients s'ils acceptent d'utiliser un appareil connecté pour suivre leur activité quotidienne (Tedesco, Barton et O'Flynn, 2017). Ces données peuvent être utilisées pour fournir de l'assistance personnalisée pour une meilleure gestion de la santé, mais aussi pour la détection précoce d'une pathologie. On peut imaginer plusieurs conséquences néfastes dues à l'inférence de ces informations, dont notamment l'augmentation de coût de l'assurance ou d'autres types de discrimination. Pour répondre aux enjeux soulevés par ces deux scénarios, nous proposons dans ce travail une solution d'assainissement des données des capteurs de mouvement de manière à masquer les attributs sensibles tout en préservant les informations d'activité contenues dans les données.

À cet effet, nous avons conçu le système *Dynamically Sanitizing Motion Sensor Data Against Sensitive Inferences through Adversarial Networks* (DYSAN) pour assainir les données des capteurs. Plus précisément, DYSAN est en mesure de construire des modèles pour assainir les données de mou-

vement afin d'empêcher les inférences sur un attribut sensible spécifié tout en maintenant un haut niveau de reconnaissance de l'activité et d'utilité pour les autres tâches d'analyse liées au suivi de l'activité (par exemple, le comptage des pas). Pour cela, DYSAN est équipé d'un module de mesure de distorsion pour limiter la modification des données originales. L'approche, qui utilise plusieurs réseaux entraînés de manière compétitive, doit donc apprendre des modèles d'assainissement afin de trouver le meilleur compromis pour faire face à ces objectifs d'optimisation contradictoires.

Notre approche vise aussi à prendre en compte l'aspect hétérogène des données collectées par les capteurs de mouvement. Ces données dépendent de la personne utilisatrice et reflètent intrinsèquement les caractères physiologiques et physionomiques de chaque personne, par exemple la manière dont elle se déplace. Ainsi, un modèle d'assainissement unique ne peut pas faire face à l'hétérogénéité des données et fournir le meilleur compromis utilité/vie privée pour toutes les personnes au fil du temps. En conséquence, DYSAN construit un ensemble de modèles d'assainissement, en explorant différentes combinaisons d'hyperparamètres qui peuvent conduisent à des équilibres différents. Ce faisant, DYSAN est capable d'évaluer les modèles assainis entraînés et de sélectionner dynamiquement le modèle offrant le meilleur compromis au fil du temps en fonction des données des capteurs entrants.

À partir de notre évaluation sur des jeux de données réels, dans lesquels le *sexe* est considéré comme l'information sensible à cacher en raison du risque possible de discrimination, nous avons pu montrer que DYSAN peut limiter considérablement l'inférence du genre jusqu'à 41%, tout en n'induisant qu'une baisse de 3% dans la reconnaissance des activités. En plus de préserver la reconnaissance d'activité, DYSAN, en limitant la distorsion des données, préserve également l'utilité des données du capteur pour d'autres tâches analytiques telles que l'estimation du nombre de pas. Nous montrons aussi que la sélection dynamique du modèle de DYSAN permet d'adapter l'assainissement en fonction des données entrantes de chaque personne. Cette sélection dynamique du modèle est particulièrement utile pour généraliser la capacité d'assainissement apprise sur un ensemble de données (utilisé pour construire les modèles d'assainissement) à un autre ensemble avec de nouvelles personnes ayant potentiellement des comportements différents. Enfin, nous avons montré que le coût d'exploitation de DYSAN sur un téléphone intelligent est compatible avec le traitement en temps réel et que la consommation d'énergie reste raisonnable.

Le projet DYSAN a été mené conjointement avec des chercheurs de l'Inria en France, Antoine Boutet, Carole Frindel et Théo Jourdan, et publié dans la conférence *ACM AsiaCCS* en *2021* : Antoine

Boutet, Carole Frindel, Sébastien Gambs, Théo Jourdan et al. (mai 2021). « DYSAN : Dynamically sanitizing motion sensor data against sensitive inferences through adversarial networks ». In : *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*. Sous la dir. de Jiannong Cao 0001, Man Ho Au, Zhiqiang Lin et Moti Yung. ACM, p. 672-686. doi : 10.1145/3433210.3453095. Avec Théo Jourdan, je suis un des auteurs principaux de cet article. En particulier, j'ai contribué à cette recherche en proposant les modèles utilisés pour l'assainissement, en faisant l'analyse des résultats de prédictions d'activité et de protection de l'attribut sensible, et enfin j'ai participé à la rédaction de l'article, notamment la description de notre architecture et le prétraitement effectué sur les données.

3.1 DYSAN : modèle système et définition du problème

DYSAN a pour objectif d'empêcher l'inférence des attributs sensibles à partir des données provenant du domaine de la santé. Dans cette section, nous présenterons la vision d'ensemble de notre système, nous poserons clairement le problème à résoudre, et enfin, nous présenterons la structure et l'utilisation de notre approche, ainsi que les différents mécanismes et les outils qui permettront la construction des modèles d'assainissement.

3.1.1 Vue d'ensemble et modèle système

Nous considérons une application mobile installée sur le téléphone intelligent d'une personne, dont le but est de mesurer son activité physique. Nous supposons dans notre cas que le téléphone est un environnement de confiance et totalement sous contrôle de la personne, tandis que le fournisseur de service responsable du traitement et de la supervision des données collectées ne l'est pas.

Les données sont collectées par une application mobile au travers de capteurs (tels que l'accéléromètre, le gyroscope et le magnétomètre). Cette dernière n'effectue aucun traitement local et transmet les informations collectées à un serveur distant (ex : un service infonuagique) qui utilise des systèmes de classification automatique pour détecter l'activité (ou toutes autres caractéristiques d'activités physiques) réalisée(s) par la personne en fonction des données de capteurs reçues. Nous considérons que le modèle d'adversaire du serveur distant est le modèle *honnête-mais-curieux*, c'est-à-dire que ce dernier essaye d'inférer des informations sensibles à partir des données observées sans toutefois dévier du protocole.

Dans le reste de cette recherche, nous considérons le genre comme étant l'attribut sensible à protéger

qui peut être déduit à partir des activités réalisées ainsi que des fréquences associées lorsque la distribution est par exemple déséquilibrée entre les hommes et les femmes. Notons toutefois que notre approche est générique et peut être appliquée dans un contexte plus large et pour d'autres attributs que ceux considérés ici.

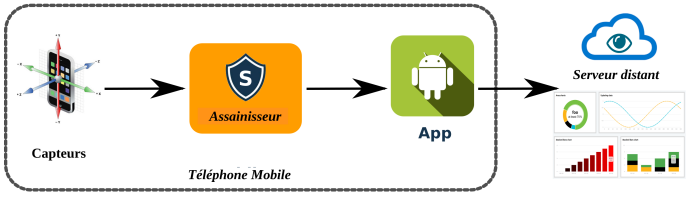


FIGURE 3.1 – DYSAN assainissement local des données de capteurs afin de prévenir l'inférence d'attribut sensible par le service infonuagique, tout en permettant la détection d'activités réalisées par les personnes participantes, ainsi que des statistiques relatives aux activités physiques.

Un aperçu de notre approche DYSAN est présenté en figure 3.1. De même que GANSan, DYSAN prévient l'exploitation non-souhaitée des données de capteurs en assainissant les informations pour empêcher l'inférence de l'information sensible tout en préservant suffisamment d'informations pour maintenir l'utilité des données. Pour ce faire, DYSAN limite la distorsion introduite entre les données brutes et celles assainies. Une fois l'assainissement réalisé localement sur le mobile, les données sont transmises au service distant pour que ce dernier réalise la détection d'activités et génère les statistiques associées.

Idéalement, l'étape d'assainissement devrait être réalisée le plus tôt possible dans le processus de traitement des informations collectées, afin de limiter l'accès aux données brutes par d'autres applications. DYSAN pourrait par exemple être déployé dans l'environnement de confiance du smartphone afin de garantir que les autres applications du téléphone ne puissent avoir accès qu'aux données assainies, sans connaissance de leurs valeurs brutes.

3.1.2 Définition du problème

Formellement, nous considérons des données A provenant de capteurs tels que l'accéléromètre et le gyroscope, qui échantillonnent un signal triaxial à une fréquence de $50Hz$. Pour réaliser la détection d'activités au fil du temps, le signal brut est divisé en fenêtres temporelles glissantes dont chacune est associée à une seule activité. La taille de la fenêtre est un paramètre difficile à choisir, particulièrement pour les tâches de reconnaissance d'activités et elle doit-être bien calibrée. En effet, une fenêtre

de petite taille diviserait le signal correspondant à une activité, tandis qu'une large fenêtre englobera les signaux de plusieurs activités différentes. Dans notre cas, nous avons fixé la taille T de la fenêtre à 2.5 secondes, avec un chevauchement de 50%, c'est-à-dire que pour deux fenêtre consécutives $n-1$ et n, la moitié supérieure de la fenêtre $n-1$ correspond à la moitié inférieure de la fenêtre n (les données de la fenêtre $n-1$ sont comprise dans l'intervalle de temps $[\frac{10n-15}{4}s, \frac{10n-5}{2}s]$ et celles de la fenêtre n dans l'intervalle $[\frac{10n-10}{4}s, \frac{10n}{2}s]$, pour tout $n > 1$). Ce choix est notamment guidé par le fait que la cadence moyenne de marche est supérieure à $1,5$ pas par seconde (BenAbdelkader, Cutler et Davis, 2002). La durée choisie pour la fenêtre choisie correspond ainsi à au temps moyen pour effectuer deux pas de marche.

Nous considérons une population de N personnes participantes dont les données sont contenues dans l'ensemble de données R. Cet ensemble de données est composé de données de capteurs, des étiquettes associées aux activités réalisées par chaque personne (dénotés par l'attribut Y), de l'attribut sensible S et d'un horodatage : $R = \{A, Y, S\}$ avec $A = (A_1, \ldots, A_T)$. L'objectif de DYSAN est de protéger les données de capteur contre l'inférence de l'attribut, tout en maintenant l'utilité des données. Nous souhaitons ainsi apprendre une collection d'assainisseurs $S_{an_{\alpha,\lambda,\beta}}$ pour diverses valeurs d'hyperparamètres α, λ et β. Parmi ces paramètres, α contrôle la protection de l'attribut sensible, λ est utilisé pour maintenir la capacité de détection des activités tandis que β aidera à la préservation de l'utilité des données pour des tâches subséquentes.

Chaque assainisseur transformera les données originales R en une version assainie $\bar{R}$ dont la prévention de S et l'utilité correspondront aux hyperparamètres choisis : $\bar{R} = S_{an_{\alpha,\lambda,\beta}}(R) = \{\bar{A}, Y, S\}$; $\bar{A} = (\bar{A}_1, \ldots, \bar{A}_T)$. L'ensemble d'assainisseurs est construit de sorte qu'il soit difficile pour un discriminateur D_{isc} de prédire S à partir des données assainies ainsi que des activités, $\{\bar{A}, Y\}$, tandis qu'un prédicteur P_{red} serait en mesure de maintenir le même niveau d'exactitude de prédiction des activités sur l'ensemble assaini que sur l'ensemble de données original. De plus, tout comme GAN-San, DYSAN se doit d'introduire le minimum de perturbations dans l'ensemble de données assaini $\bar{R}$, ce qui favorise la préservation de l'utilité.

DYSAN a pour but d'adapter dynamiquement les hyperparamètres au fil du temps et des données entrantes de chaque personne. En effet, bien qu'un modèle entraîné puisse fournir en moyenne le meilleur compromis utilité/vie privée pour toutes les personnes de l'ensemble d'entraînement, ce modèle peut varier lors du test sur de nouvelles données de personnes et ne pas fournir de bons

compromis pour certains (malgré le fait qu'en moyenne, le modèle se comporte bien sur tous). Ceci est particulièrement possible si les données d'une nouvelle personne participante ne suivent pas la distribution de l'ensemble d'entraînement. De manière plus formelle, l'objectif est de pouvoir associer à chaque fenêtre de données, l'assainisseur $\widehat{San_{\alpha,\lambda,\beta}}$ fournissant le meilleur compromis utilité/vie privée pour ladite fenêtre. Ce compromis est défini par une métrique tenant compte de l'exactitude de prédiction de l'activité et la protection de l'attribut sensible.

3.2 DYSAN : Assainisseur dynamique

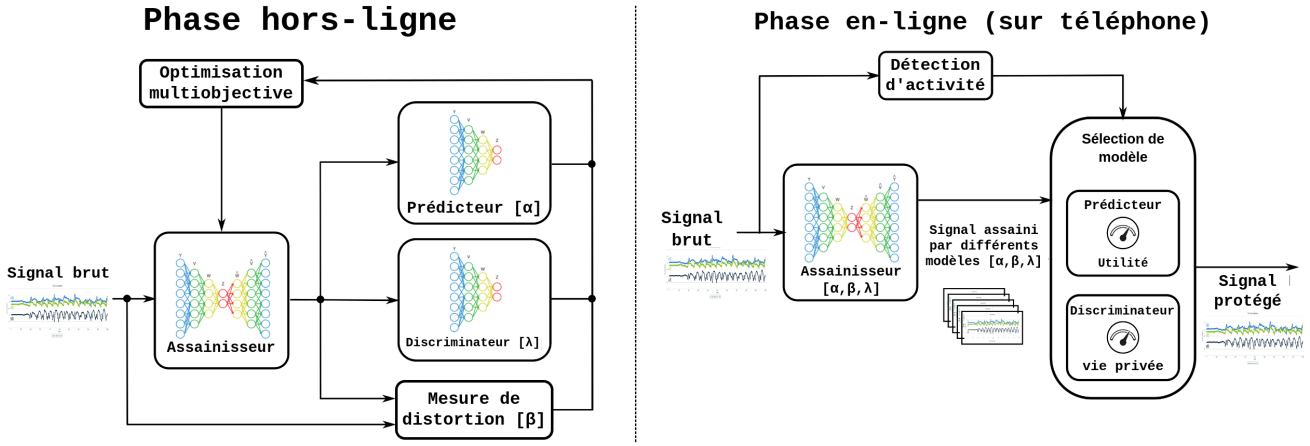


FIGURE 3.2 – DYSAN est composé de deux étapes : une phase d'entraînement hors-ligne (à gauche) et une étape en ligne (à droite). La phase d'entraînement est réalisée une seule fois, et a pour but la construction des différents modèles d'assainisseur en fonction des différentes combinaisons d'hyperparamètres. Une fois les modèles déployés sur le téléphone, la phase en ligne a pour but de choisir dynamiquement le meilleur modèle parmi la collection d'assainisseurs pour chaque lot de données entrantes.

Avant son déploiement sur le téléphone intelligent d'une personne, une collection d'assainisseurs (correspondant chacun à différents niveaux de compromis utilité/vie privée) est construite. Cette étape correspond à la phase hors-ligne de DYSAN. Une fois les modèles déployés, DYSAN entre dans sa phase en ligne qui consiste en la sélection dynamique du modèle affichant le meilleur compromis pour le lot de données reçu. Ces deux phases sont résumées dans la figure 3.2.

3.2.1 Construction des multiples assainisseurs

La phase d'entraînement hors-ligne est réalisée une seule fois et a pour but de construire (c'est-à-dire d'entraîner) plusieurs modèles d'assainisseurs. L'entraînement est réalisé avec un ensemble de données de référence utilisé en reconnaissance d'activités, *MotionSense* que nous décrirons en section 3.3.1. DYSAN est composé de plusieurs blocs : un assainisseur, un discriminateur, un prédicteur, une mesure de distorsion et une fonction de perte multiobjective, comme montré en figure 3.2.

— *Discriminateur.* Le discriminateur D_{isc} guide l'assainisseur dans le processus de retrait d'informations liées à l'attribut sensible $S \in \{0, 1\}$. En pratique, nous utilisons des réseaux de neurones convolutionels, qui pour des séries temporelles, sont adaptés à la capture de caractéristiques invariantes dans le temps (Ismail Fawaz, Forestier, Weber, Idoumghar et al., 2019). Le discriminateur est entraîné à minimiser le taux d'erreur équilibré (*cf.* chapitre 1, section 1.4), ce qui le rendrait ainsi apte à inférer S. Nous dénoterons cette fonction de perte par $Loss_{Sensitive}$:

$$Loss_{Sensitive} = BER(D_{isc}(\bar{A}, Y), S) = \frac{1}{2}(\sum_{s=0}^{1} P(D_{isc}(\bar{A}, Y) \neq s | S = s)). \tag{3.1}$$

— *Prédicteur.* Le prédicteur P_{red} est utilisé pour aider l'assainisseur à préserver le plus d'informations possibles concernant la reconnaissance d'activités. Pour cela, le prédicteur est entraîné à prédire l'activité de la personne à partir des données assainies. L'objectif est de pouvoir maximiser l'exactitude de prédiction de cette tâche, et par ricochet de minimiser le taux d'erreur équilibré ($BER(P_{red}(\bar{A}), y)$) utilisé comme fonction de perte pour ce modèle. Tout comme le discriminateur, le prédicteur est aussi un réseau de neurones convolutionels. La fonction de perte du prédicteur sera nommée $Loss_{Activities}$.

— *Mesure de distorsion.* La dernière contrainte de l'assainisseur porte sur la minimisation de la distorsion entre les données brutes et celles assainies. L'objectif ici étant de minimiser les modifications introduites dans les données afin de permettre l'utilisation de l'ensemble de données pour des tâches subséquentes. La mesure de distorsion sera quantifiée au travers de la fonction de perte MAE ou L_1 appliquée sur chacun des attributs les uns indépendamment des autres. Pour deux vecteurs A_i et $\bar{A}_i$, correspondant respectivement à la version originale (brute) et assainie des données, la fonction est définie comme étant (équation 3.2) :

$$L_1(A_i, \bar{A}_i) = \frac{1}{N_A} \sum_{j=1}^{N_A} |a_{ij} - \bar{a}_{ij}|, \tag{3.2}$$

où N_A est le nombre de valeurs possibles pour un attribut (ex : le nombre d'axes de l'accéléromètre ou du gyroscope), $a_{ij} \in A_i$ et i dénote une observation unique dans la fenêtre de largeur T.

— *Assainisseur.* L'assainisseur S_{an} modifie les données brutes reçues en entrée afin de cacher l'information privée, tout en maintenant l'utilité de l'ensemble de données. L'assainisseur ici peut être assimilé à un auto-encodeur. Il prend en considération le retour du discriminateur, du prédicteur et de la mesure de distorsion pour produire une version assainie des données. Ces différents retours sont intégrés dans notre fonction multiobjective que l'assainisseur est entraîné à minimiser.

— **Fonction multi-objectif.** L'assainisseur est optimisé par la fonction $J^{S_{an}}$, qui oriente les transformations de l'ensemble original en sa version assainie. La fonction prend en considération (1) la capacité de détection de l'activité, évaluée par la sortie du prédicteur, (2) la capacité de détection de l'attribut sensible (mesurée par le discriminateur) et (3) la quantité de dommage introduit dans les données :

$$\begin{aligned} J^{S_{an}}(R,\, S_{an},\, D_{isc},\, P_{red}) = \{&\alpha * d_s(S, D_{isc}(S_{an}(R))), \\ &\lambda * d_p(Y, P_{red}(S_{an}(R))), \\ &\beta * d_r(R, S_{an}(R))\}, \end{aligned} \tag{3.3}$$

$d_s(x) = \frac{1}{2} - Loss_{Sensitive}$, $d_p = Loss_{Activities}$, $d_r = \{l1(a_{:,j} \text{ et } \bar{a}_{:,j}), \ldots\}$ avec $a_{:,j}$ représentant la dimension de tous les pas de temps dans une fenêtre glissante. Pour rappel, la protection mesurée par le BER est maximale à la valeur $\frac{1}{2}$. En conséquence, le terme $\frac{1}{2}$ dans l'équation 3.3 est utilisé pour amener l'assainisseur à maximiser l'erreur de prédiction de S par le discriminateur. Pour rappel, le paramètre α représente l'importance relative de la protection de la vie privée tandis que λ contrôle l'utilité des données (c'est-à-dire la qualité de détection des activités). Pour faciliter la recherche d'hyperparamètres, nous imposons : $\alpha + \lambda + \beta = 1$. De ce fait, seuls les hyperparamètres α et λ doivent être choisis puisque $\beta = 1 - (\alpha + \lambda)$.

3.2.2 Phase d'entraînement

Durant la phase d'entraînement, nous construisons un assainisseur par combinaison d'hyperparamètres α et λ, ceci afin d'explorer le domaine de la fonction multiobjective. Cette exploration permettra à DYSAN de choisir le meilleur modèle pour chaque personne durant la phase en ligne. La procédure d'entraînement est résumée dans l'algorithme 2.

Algorithm 2 DYSAN algorithme d'entraînement

1: **Intrants :** R, λ, α, max_epoch, $batch\_size$, K_{pred}, K_{disc}.

2: **Extrants :** S_{an}, D_{isc}, P_{red}.

3: **train(M, trParams) :** entraînement du modèle M en utilisant les paramètres trParams.

4: **freeze(M) :** Fixation des paramètres du modèle M pour éviter toute modification.

▷ Initialisation

5: S_{an}, D_{isc}, P_{red}, $R_d = \mathsf{shuffle}(R)$, $R_p = \mathsf{shuffle}(R)$

6: $\mathsf{Iterations} = \frac{|D|}{\mathsf{batch\_size}}$

▷ Procédure d'entraînement

7: **for** $e = 1$ **to** max_epoch **do**

8: **for** $i = 1$ **to** Iterations **do**

9: Prélever un lot B de taille batch_size à partir de l'ensemble R

10: train(S_{an}, B, $J^{S_{an}}$, α, λ, freeze(P_{red}), freeze(D_{isc}))

11: **for** $k = 1$ **to** K_{pred} **do**

12: Prélever un lot B de taille batch_size à partir de R_p

13: train(P_{red}, B, $Loss_{Activities}$, freeze(S_{an}))

14: **end for**

15: **for** $k = 1$ **to** K_{disc} **do**

16: Prélever un sous-ensemble B de taille batch_size à partir de R_d

17: train(D_{isc}, B, $Loss_{Sensitive}$, freeze(S_{an}))

18: **end for**

19: **end for**

20: **end for**

Afin d'optimiser le compromis utilité/vie privée pour chaque combinaison de α et λ (ligne 1, algorithme 2), les trois réseaux de neurones sont entraînés de manière compétitive, tout comme dans GANSan, entre l'assainisseur et le discriminateur. Plus précisément, l'assainisseur est entraîné de sorte que ses extrants soient en mesure de tromper le discriminateur, tout en restant utilisable pour le prédicteur et lui permettant de maintenir le plus possible l'exactitude de prédiction originale.

L'entraînement de DYSAN est réalisé en alternant l'entraînement de chaque modèle jusqu'à convergence, ou jusqu'à ce qu'un nombre maximum d'époques soit atteint. Spécifiquement, à la suite de l'initialisation (lignes $1-8$), l'entraînement de l'assainisseur commence avec $J^{S_{an}}$, tandis que le pré-

dicteur et le discriminateur sont figés dans leurs paramètres (lignes $11-12$). Une fois l'entraînement de l'assainisseur réalisé pour un certain nombre d'itérations, le prédicteur et le discriminateur sont entraînés parallèlement avec leur fonction de pertes respectives (lignes $13-20$). Ces deux modèles sont entraînés jusqu'à la convergence, c'est-à-dire que la fonction de perte ne décroît plus ou jusqu'à un nombre maximum d'itérations, qui sont K_{pred} pour le prédicteur et K_{disc} pour le discriminateur.

3.2.3 Phase de déploiement en ligne

Une fois déployé sur le téléphone, DYSAN se compose de quatre éléments (*cf.* figure 3.2) : l'assainisseur, le discriminateur, le prédicteur et le module de détection d'activité. Spécifiquement, DYSAN connait toutes les versions de l'assainisseur, du prédicteur et du discriminateur construits durant la phase d'entraînement. Puisque chaque ensemble de trois modèles sont obtenus par une combinaison d'hyperparamètres représentant différents compromis utilité/vie privée, la sélection du modèle pour chaque lot de données d'entrée est réalisée au travers de la fonction $S(P,U) = xU + yP$, où U correspond à l'évaluation de la détection d'activités réalisée par le prédicteur, et P une métrique de vie privée obtenue par le discriminateur (qui produit la probabilité p d'appartenance à un groupe $S = 1$), telle que $P = 1 - |0,5 - p|$. Cette métrique est maximale lorsque le discriminateur est incapable de prédire l'attribut sensible, et minimale lorsqu'il le prédit parfaitement. Enfin, x et y correspondent à des coefficients positifs choisis tels que $x + y = 1$. Ils permettent à la personne finale de mieux paramétrer le choix de l'assainisseur en fonction de ses préférences entre utilité et vie privée.

Pour trouver le meilleur assainisseur au fil du temps et des données reçues (et en fonction des coefficients x et y), DYSAN évalue le compromis utilité/vie privée de tous les modèles. Cette sélection nécessite notamment la connaissance de l'activité réalisée et l'attribut sensible. Bien que l'attribut sensible puisse être donné par la personne utilisatrice de manière asynchrone, les données de capteurs ne sont pas étiquetées. En effet, le système est supposé déployé dans l'environnement protégé du téléphone, la personne utilisatrice n'y a donc pas forcément accès. De plus, ce dernier ne peut étiqueter toutes les données pendant qu'il réalise l'activité physique (traitement en temps réel).

Nous utilisons en conséquence le module de détection d'activité pour annoter certaines données sur le téléphone. Plus précisément, nous demandons à la personne participante de suivre une procédure de calibration spécifique à l'installation de l'application DYSAN. Durant cette procédure, il lui est demandé de réaliser une série d'activités différentes pour une courte période afin d'entraîner un

prédicteur spécifique pour la détection de ses activités. Étant donné la petite quantité de données disponibles pour l'entraînement de ce classifieur, nous avons utilisé une version modifiée des forêts aléatoires (*Random Forest ou RF*) adaptée à cette tâche (Jourdan, Boutet et Frindel, 2018). Le classifieur RF est ensuite utilisé pour étiqueter les données brutes afin d'évaluer l'utilité de tous les assainisseurs. Cette évaluation est réalisée sur une base régulière avec l'exactitude de prédiction moyenne qui est calculée sur cette base. En suivant cette procédure, DYSAN est en mesure d'identifier au fil du temps et des données arrivantes, l'assainisseur qui fournit le meilleur compromis utilité/vie privée.

Le code de DYSAN est accessible à l'adresse suivante : `https://github.com/DynamicSanitizer/DySan`

3.3 DYSAN : cadre expérimental

Dans cette section, nous décrivons le protocole d'évaluation de notre approche, qui contiendra une description des ensembles de données spécifiques à notre contexte d'étude, les métriques d'évaluation de protection et d'utilité et la méthodologie d'évaluation. Nous y présenterons aussi les principales approches concurrentes de la nôtre.

3.3.1 Ensemble de données

Pour évaluer DYSAN, nous avons utilisé deux ensembles de données publiquement disponibles et qui sont beaucoup utilisés dans la littérature pour la reconnaissance d'activités : MotionSense et MobiAct. Ces ensembles sont composés de données de capteurs de personnes effectuant des activités cyclostationnaires, c'est-à-dire des activités pouvant être modélisées par des déplacements en pas.

— *MotionSense* (Reyes-Ortiz, 2015) contient les données d'accéléromètre (accélération et gravité) et gyroscopiques capturées à une fréquence constante de $50Hz$, sur un iPhone 6s maintenu dans la poche avant des participants(es). Vingt-quatre participants y réalisent six activités différentes (descendre des marches, monter des escaliers, marcher, jogger, s'asseoir et se lever).

— *MobiAct* (Vavoulas, Chatzaki, Malliotakis, Pediaditis et al., 2016) est composé des données de capteurs de 58 personnes participantes suivis sur plus de 2500 essais. Les données sont capturées avec un smartphone (Samsung Galaxy S3) maintenu en poche. Les signaux pour cet ensemble de données sont aussi mesurés par l'accéléromètre et le gyroscope. Les

personnes participantes réalisent neuf activités courantes de tous les jours. Cependant, pour notre expérience, nous nous sommes limités uniquement aux activités identiques à celle de MotionSense.

Ces deux ensembles sont équilibrés, c'est-à-dire qu'ils contiennent autant d'hommes que de femmes. L'activité de marche (*walking*) est la plus représentée comparativement aux autres (*cf.* section 3.4.3). Chaque activité est réalisée un nombre égal de fois par toutes les personnes dans les deux ensembles. Par conséquent, la corrélation entre l'attribut sensible et l'activité est minimale. Puisque ces ensembles partagent des caractéristiques similaires, nous pouvons mesurer la *transférabilité* de nos modèles d'un jeu de données à un autre, c'est-à-dire que les modèles entraînés sur un des ensembles peuvent être utilisés pour assainir les données de l'autre. Ce mode d'évaluation n'a pas encore été utilisé dans la littérature sur l'assainissement des données de capteurs. Pour nos expériences, l'ensemble de données est divisé en ensemble d'entraînement et de test, suivant les proportions 2/3 des données utilisées pour l'entraînement et 1/3 pour le test.

3.3.2 Méthodes de l'état-de-l'art

Nous comparons les performances de DYSAN à plusieurs méthodes de l'état de l'art, dont le classifieur RF modifié (Jourdan, Boutet et Frindel, 2018) et d'autres basées sur les RAG (Malekzadeh, Clegg, Cavallaro et Haddadi, 2018, 2019 ; Raval, Machanavajjhala et Pan, 2019). Concernant ces dernières, les auteurs utilisent des architectures de réseaux légèrement différentes de la nôtre. Pour réaliser une comparaison équitable, nous proposons d'implémenter leurs fonctionnalités dans notre architecture (nombre de couches, type de CNN, ...). Cette méthodologie permet d'évaluer les caractéristiques principales adoptées dans ces références, sans influence des artéfacts liés aux structures des réseaux.

— *ORF.* L'approche ORF est proposée dans l'article Jourdan, Boutet et Frindel (2018). Pour limiter leur exposition, les données brutes sont prétraitées sur le téléphone de la personne et uniquement les caractéristiques pertinentes à une tâche sont transmises au service d'infonuagique. Ces caractéristiques spécifiques à la tâche (par exemple la reconnaissance d'activités) sont choisies soit dans le domaine temporel, soit dans le domaine fréquentiel. L'approche a été proposée originellement pour prévenir la réidentification des personnes, mais nous l'avons adaptée dans notre contexte à la protection de l'attribut sensible. Plus précisément, nous détectons dans un premier temps les caractéristiques les plus corrélées au

genre, puis ces caractéristiques sont normalisées dans le domaine fréquentiel, et celles non pertinentes du domaine temporel pour la classification sont retirées.

— *GEN.* De même que DYSAN, GEN (*Guardian Estimator Neutralizer*) s'appuie sur une architecture basée des modèles adversariaux pour optimiser l'utilité et la vie privée. Leur architecture diffère notamment de la nôtre par le fait que leur classifieur utilisé n'est qu'entraîné une seule fois sur les données originales, pour identifier aussi bien l'attribut sensible que l'activité. De même, leur auto-encodeur est aussi entraîné une seule fois, sans tenir compte des distorsions introduites dans les données. Finalement, le modèle utilisé dans la phase en ligne est unique pour toutes les personnes, et correspond à celui ayant obtenu les meilleures performances durant la phase d'entraînement. GEN a été implémenté suivant notre structure de modèles. Son évaluation a été menée dans le contexte de transfert d'apprentissage, et nous avons aussi utilisé les modèles originaux entraînés sur MotionSense [1] pour mesurer les performances sur MobiAct.

— *Olympus.* L'approche Olympus (Raval, Machanavajjhala et Pan, 2019) est similaire à GEN. Toutefois, celle-ci utilise deux réseaux différents, un premier pour la prédiction de l'attribut sensible et un second pour la reconnaissance d'activités alors que GEN utilise un seul réseau pour ces deux tâches. En outre, les modèles sont ici entraînés par une méthode itérative similaire à la nôtre. Cependant, la fonction de coût ne tient pas compte des distorsions introduites dans les données, en plus du fait qu'un seul modèle est utilisé pour toutes les personnes. Cette approche, tout comme ORF, est utilisée pour empêcher la réidentification des personnes. Nous l'avons donc adaptée à notre contexte en utilisant notre architecture.

— *MSDA.* MSDA (Malekzadeh, Clegg, Cavallaro et Haddadi, 2019) peut être considérée comme une évolution de l'approche Olympus, dans laquelle la distorsion des données est prise en compte. De même que toutes les autres approches précédentes, MSDA utilise un unique modèle durant la phase en ligne. Nous avons aussi adapté MSDA pour la détection d'activités en utilisant notre architecture. MSDA est l'approche de référence la plus proche de DYSAN, sans toutefois la sélection dynamique et l'entraînement multiple de modèles.

Les résultats que nous reportons sont obtenus en fixant les paramètres x et y aux valeurs respectives 0.1 et 0.9. Pour rappel, ces paramètres permettent de sélectionner le modèle permettant de fournir le meilleur compromis utilité/vie privée lors de la sélection dynamique de modèles, et reflètent les

1. `https://github.com/mmalekzadeh/motion-sense/tree/master/codes/gen_paper_codes`

préférences de la personne (*cf.* section 3.2.3). L'impact de différentes valeurs de x et y ont été évalués et présentés en annexe D dans notre article (Boutet, Frindel, Gambs, Jourdan et al., 2021). On peut y observer que le paramètre de protection a une influence plus importante sur les résultats que celui de l'utilité, notamment à cause du fait que l'utilité mesurée peut rester assez élevée et proche du maximum possible (atteignable lorsque la protection n'est pas prise en compte, $y = 0$) lorsque le paramètre de protection prend la valeur maximale (perte d'utilité inférieure à 4%).

3.3.3 Métriques d'évaluation

Nous évaluons DYSAN en utilisant des métriques d'utilité et de protection de la vie privée, en plus de quelques-unes liées au déploiement du modèle.

— *Utilité.* Dans notre contexte de suivi d'activités physiques, la première métrique d'utilité considérée est l'exactitude de prédiction d'un classifieur entraîné pour la reconnaissance d'activités. En d'autres termes, nous mesurons le nombre d'activités correctement identifiées par le modèle à partir des prédictions réalisées. L'idéal pour cette métrique serait la valeur de 1, qui indique que le modèle n'a commis aucune erreur d'identification. En plus de cette reconnaissance d'activités, nous calculons sur l'ensemble de données assainies d'autres caractéristiques d'activités physiques telles que le nombre de pas, que nous comparerons avec celles que nous pouvons obtenir sur l'ensemble de données brutes. Pour cela, les données originales et assainies sont normalisées pour les maintenir dans le même intervalle de valeurs et faciliter la comparaison. Ensuite, nous calculons le seuil du pic d'accélération (*Peak Acceleration Threshold*) (Abadleh, Al-Hawari, Alkafaween et Al-Sawalqah, 2017). Ce seuil est utilisé pour détecter les différentes foulées des personnes, l'hypothèse étant que les pics d'accélérations représentent le mouvement vertical correspondant ou mouvement de levée de jambe. Plus précisément, nous utilisons la méthode issue de la recherche "Adaptiv : An Adaptive Jerk Pace Buffer Step Detection Algorithm" (`https://github.com/danielmurray/adaptiv`) pour l'estimation du nombre de pas. Cette méthode consiste en une détection de foulées de personnes en se basant sur une mise à jour dynamique du seuil d'accélération chaque fois qu'un nouveau pas est détecté, permettant ainsi d'éviter les faux pics d'accélération correspondant à d'autres mouvements que des foulées.

— *Protection de la vie privée.* La mesure de la protection des informations personnelles se fait au travers du $SAcc$, qui est l'exactitude de prédiction de l'attribut sensible. Dans notre cas,

l'attribut sensible est le genre des personnes. Pour nos jeux de données qui sont équilibrés, une exactitude de prédiction de $0,5$ serait une valeur idéale, car cela correspond à des prédictions aléatoires.

— *Métriques système.* Enfin, nous devons évaluer les surcoûts liés à l'utilisation de notre application localement sur un smartphone. Pour cela, nous mesurons l'utilisation CPU pour l'assainissement des données brutes et la consommation énergétique due à l'utilisation temps réel de DYSAN.

3.3.4 Méthodologie

DYSAN est uniquement entraîné sur l'ensemble de données MotionSense, nos résultats reportés pour MobiAct sont obtenus par transfert d'apprentissage, en utilisant les modèles entraînés sur MotionSense. Durant la phase d'entraînement, nous évaluons des valeurs entre $0,1$ et $0,9$ avec un pas de $0,1$ pour α et λ ce qui correspond à 36 modèles d'assainissement différents. Ces modèles sont entraînés sur 300 époques et la taille de chaque lot de données est fixée à 256 échantillons. Le nombre d'époques a été choisi de sorte que les modèles convergent et que les performances des modèles ne montrent pas d'améliorations significatives dans les époques subséquentes.

Dans la phase en ligne, nous choisissons de mettre l'accent sur la protection de l'attribut sensible. En conséquence, puisque ce choix est contrôlé par les coefficients x pour l'utilité et y pour la vie privée (section 3.2.3), nous les avons fixés aux valeurs $0,1$ et $0,9$. Concernant le détecteur d'activité de type forêt aléatoire utilisé dans la phase en ligne, ce dernier utilise un vecteur de caractéristiques (ou descripteurs) extraites du signal brut. Ces descripteurs ont été choisis sur la base d'un article de recherche concernant le choix effectif de caractéristiques pour la reconnaissance de démarche (Sprager et Juric, 2015). Enfin, nous utilisons une validation croisée de 4-blocs, dans lequel l'ensemble de données est divisé aléatoirement en quatre sous-ensembles de taille équivalente.

Les résultats reportés correspondent à la moyenne obtenue sur 10 répétitions de chaque expérience réalisée sur une plateforme de calcul. Nous évaluerons DYSAN suivant cinq axes principaux :

— *Évaluation de la détection des activités et de la protection de l'attribut sensible.* La capacité de notre application à inférer le genre et les activités réalisées par la personne participante à partir des données assainies sera évaluée et nos performances seront comparées à plusieurs classifieurs qui pourraient être utilisés par le service distant, notamment Gradient Boosting (GB), Perceptron Multicouche (MPL), des réseaux récurrents de type LSTM, les arbres de

décision (DT), les forêts aléatoires (RF) et la régression logistique (LR).

— *Impact des perturbations introduites dans les données.* L'utilité de l'ensemble de données assaini ne se limite pas juste à la détection d'activités, mais aussi à la conservation d'informations plus subtiles. L'objectif de cette analyse est de montrer la capacité de DYSAN à maintenir le plus d'utilité dans les données de sorte qu'elles soient utilisables pour d'autres tâches. Pour cela, nous nous appuyons sur le comptage du nombre de pas effectué par le participant à partir des signaux de l'ensemble MotionSense. De plus, nous mesurons aussi la déformation temporelle dynamique (*Dynamic Time Warping* ou *DTW*) (Berndt et Clifford, 1994) entre le signal original et sa version assainie obtenue à partir de notre approche et aussi par les références. Cette métrique mesure la distorsion entre deux signaux temporels. L'objectif pour cette mesure est d'obtenir une faible valeur qui indiquerait la similarité des signaux.

— *Comparaison avec l'état-de-l'art.* Pour réaliser cette comparaison, nous nous appuyons sur le prédicteur et le discriminateur utilisé lors de l'entraînement de chacune des approches. De plus, nous considérons aussi deux versions différentes de DYSAN : une version dans laquelle les activités sont connues d'avance et DYSAN(o) dans laquelle les activités sont inférées par le classifieur forêt aléatoire (RF). La première version est rajoutée pour une comparaison équitable avec l'état de l'art, puisque notre protocole d'utilisation est différent des leurs.

— *Avantage de l'assainissement dynamique par rapport à d'autres approches.* Nous mesurons l'avantage de l'adaptation dynamique du modèle d'assainissement en fonction des données entrantes de chaque personne participante par rapport à deux approches statiques de référence : la première dans laquelle le modèle de prédiction du genre et de reconnaissance d'activités est fixe pour toutes les personnes, et la seconde dans laquelle la solution est personnalisée pour chaque profil de personne. Le premier cas représente le comportement de toutes les approches de référence, dans lesquelles le modèle considéré est celui qui fournit les meilleures performances (c'est-à-dire le compromis entre utilité et vie privée) en moyenne pour toutes les personnes participantes. Dans le second cas, le modèle d'assainissement choisi pour chaque personne participante est celui qui fournit le plus petit SAcc en termes d'inférence du genre et la meilleure reconnaissance d'activités parmi l'ensemble des modèles testés sur cette personne. Ce modèle reste fixe, c'est-à-dire qu'il n'est pas modifié en fonction de l'évolution des données entrantes (et des changements associés en termes d'activités réalisées).

— *Coût opérationnel de DYSAN sur un téléphone mobile.* Nous évaluons ici le coût de fonctionnement de DYSAN sur un téléphone intelligent, c'est-à-dire la surcharge causée par la

sauvegarde des modèles sur le téléphone, et l'énergie consommée par l'exécution de DYSAN. En effet, DYSAN évalue plusieurs modèles d'assainissement (selon chaque combinaison d'hyperparamètres α et λ exploré) avant de sélectionner celui qui produit le meilleur compromis entre utilité et protection. Les coûts associés à l'assainissement des données brutes dépendent ainsi du nombre de modèles considérés. Par contre, nous ne tenons pas compte du coût de l'apprentissage lui-même, car il s'agit d'une opération ponctuelle.

3.4 DYSAN : résultats

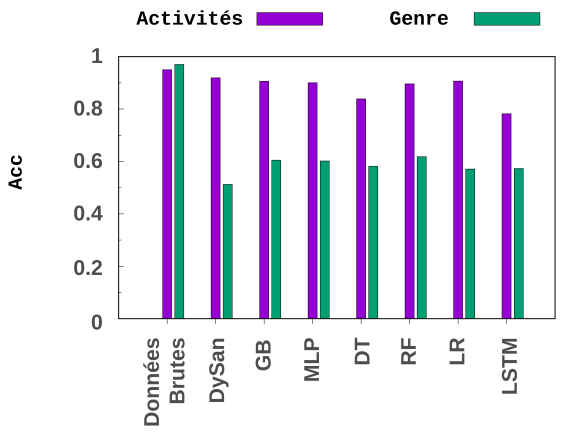
Comme mentionné dans la section précédentes, nous reportons les résultats obtenus pour l'évaluation de DYSAN en mettant en avant l'utilité et la vie privée en section 3.4.1, les perturbations introduites dans l'ensemble assaini (section 3.4.2), la comparaison avec l'état de l'art (section 3.4.3), et enfin l'avantage du choix dynamique de modèles pour l'assainissement (section 3.4.4) ainsi que le coût opérationnel de DYSAN sur un mobile (section 3.4.5).

3.4.1 Compromis utilité et vie privée

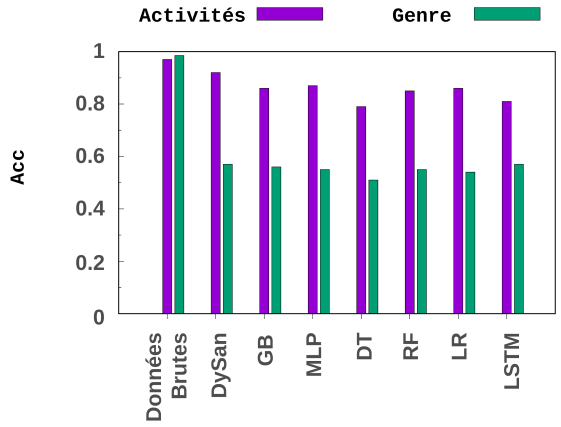
Sur la figure 3.3, nous présentons l'exactitude de prédiction du genre et de l'activité pour les deux ensembles de données, sur leur version originale et lorsqu'elles sont assainies par notre approche. Premièrement, nos résultats montrent que sans aucune protection, les services distants sont en mesure d'inférer le genre, à $98,5\%$. La détection d'activité se fait aussi très efficacement, avec une moyenne de 97%. Ensuite, nous pouvons observer que DYSAN réduit avec succès les bris de vie privée due à l'inférence de l'attribut sensible, tout en limitant la perte en détection d'activités. En effet, avec les données assainies, le service analytique distant est en mesure d'inférer le genre à hauteur de 61% et de 57% pour respectivement les ensembles MotionSense et MobiAct. En termes d'utilité, dépendemment du classifieur utilisé, l'exactitude de prédiction de l'activité varie entre 78% et 92%. Ceci représente une légère baisse comparativement à l'utilisation des données brutes. Notons toutefois que le réseau récurrent LSTM, qui est l'architecture couramment utilisée pour les séries temporelles, n'a pas de bonnes performances, contrairement à nos attentes.

3.4.2 Distorsion du signal assaini

De prime abord, notons que ORF n'est pas analysée pour cette partie, puisque celui-ci nécessite uniquement l'extraction de caractéristiques et ne conserve pas le signal une fois ces caractéristiques extraites. Des analyses subséquentes ne peuvent donc pas être réalisées.



(a) MotionSense



(b) MobiAct

FIGURE 3.3 – Les données assainies par DYSAN réduisent radicalement les risques de vie privée, comparativement aux données brutes, tout en limitant la perte en détection d'activités, quel que soit le mécanisme de classification utilisé.

Dans le tableau 3.1 nous pouvons voir qu'à partir des données assainies avec DYSAN, il existe une erreur d'approximativement 7% dans le comptage du nombre de pas entre ces données et leurs versions originales. Comparativement à notre approche, les différentes références produisent des versions assainies qui sont beaucoup plus bruitées, ce qui affecte considérablement le comptage de pas. En effet, on obtient une erreur de 74% avec Olympus, de 29% avec MSDA et de plus de 12% avec GEN. Le résultat de GEN est notable, parce que cette approche ne prend en compte la quantité de modifications introduites et n'inclue pas de contrainte de minimisation de cette mesure. Toutefois, le faible taux d'erreur s'accompagne aussi d'une faible protection de notre attribut sensible

(figure 3.4a).

À partir de MSDA - qui est l'approche la plus proche de la nôtre, on peut constater que l'utilisation d'un unique modèle pour la protection de tous les personnes n'est pas aussi efficace qu'une sélection dynamique comme assurée par notre approche, et ce, quelle que soit l'activité réalisée.

	Nombre de pas	DTW
Données brutes	14387	-
DYSAN	15321 (**+6,49** %)	12,96
GEN	12817 (**-12,25**%)	14,28
Olympus	23658 (**+64,44**%)	156,03
MSDA	18624 (**+29,45**%)	23,37

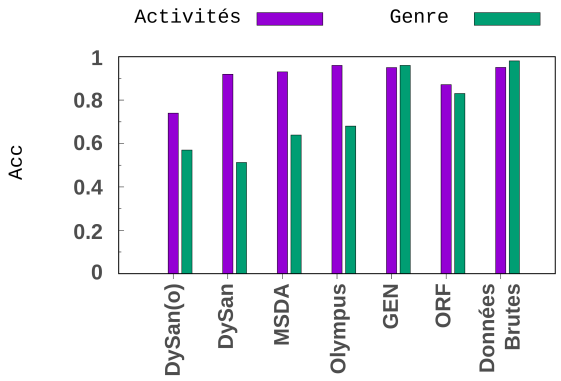
TABLEAU 3.1 – Le signal assaini obtenu avec DYSAN apparaît moins perturbé et adéquat aux tâches subséquentes, contrairement aux autres approches de l'état de l'art.

Concernant la déformation temporelle dynamique (DTW), on observe dans la table 3.1 que le signal assaini par DYSAN est celui étant le plus proche du signal original parmi toutes les approches évaluées.

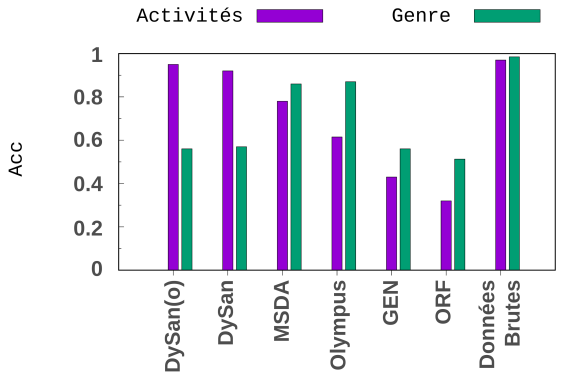
3.4.3 Analyse comparative

Sur MotionSense (figure 3.4a), l'amélioration de la vie privée avec DYSAN se fait avec un coût léger en utilité : l'inférence de l'attribut est réduite à 51%, tandis que la reconnaissance d'activité est de 92%. Pour notre version en ligne qui n'a pas connaissance des vraies annotations, les performances sont un peu plus faibles (inférence du genre à 57% et exactitude de prédiction de 75%). Cette décroissance en utilité est notamment due aux erreurs de prédiction du classifieur forêt aléatoire utilisé dans la phase en ligne pour étiqueter les données, et ainsi choisir le meilleur modèle d'assainisseur (section 3.2.3). En effet, après l'étape de calibration du classifieur sur le téléphone, les performances sont respectivement de 96% et 94% sur les ensembles MotionSense et MobiAct. En conséquence, les activités incorrectement annotées par ce classifieur conduisent à une sélection incorrecte de l'assainisseur à utiliser.

En figure 3.4b, nous observons les performances sur MobiAct. Entre autres, nous pouvons voir que DYSAN et DYSAN(o) obtiennent de meilleures performances que toutes les autres, en réduisant



(a) MotionSense



(b) MobiAct

FIGURE 3.4 – DYSAN fournit le meilleur compromis pour la protection de l'attribut sensible, comparativement aux approches de l'état de l'art, pour un coût léger en utilité.

l'inférence de l'attribut à 55% et 54% respectivement, tout en assumant une perte sur la reconnaissance d'activité de 2% et 5%. GEN et ORF réduisent tout aussi bien la possibilité d'inférer l'attribut sensible S. Cependant, cette amélioration de la protection vient avec un coût considérable en utilité, une perte de 43% et 32% dans la reconnaissance d'activités.

Nous présentons dans le tableau 3.2 les vrais positifs et faux positifs obtenus avec DYSAN sur MobiSense. Ce tableau rapporte aussi la proportion de chaque activité dans l'ensemble de données. Nous pouvons y observer que la prédiction des activités n'est pas uniforme, fort probablement en conséquence du déséquilibre de proportions de ces activités. Spécifiquement, l'activité marche (*walking*) obtient la précision la plus haute, mais elle est aussi majoritairement présente dans l'ensemble de

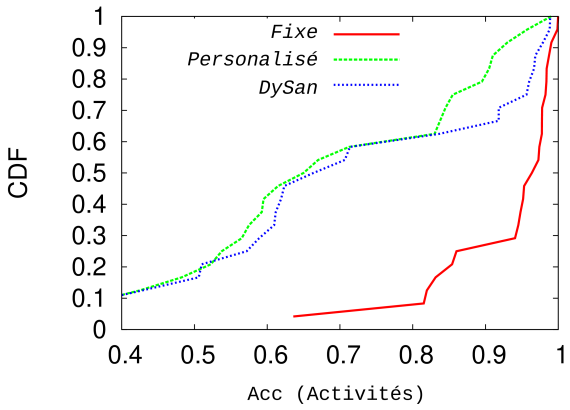
	VP	FP	Précision	Pourcentage des données
Monter des marches	221	112	66,4	17,2
Descendre des marches	223	198	53,0	20,5
Marcher	918	74	92,5	44,9
Course	216	212	50,5	17,4

TABLEAU 3.2 – Vrais Positifs (VP), Faux Positifs (FP), précision obtenus avec DYSAN et pourcentage des données de chaque activité (ensemble de données MotionSense).

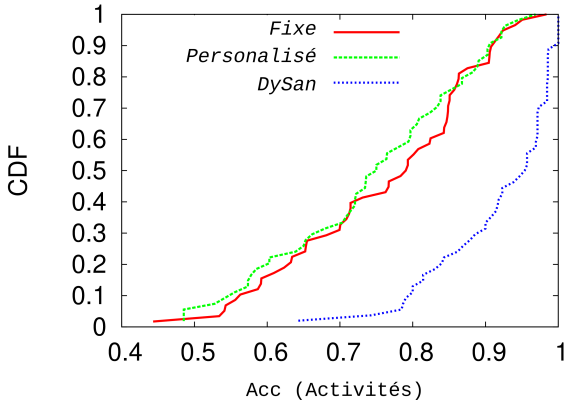
données. Les autres, moins présentes, obtiennent par conséquent de moins bonnes performances. Ces résultats peuvent aussi être dus à la taille de la fenêtre glissante utilisée dans notre approche, qui a été choisie notamment en fonction de références obtenues sur l'activité de marche (partie 3.1).

Les résultats montrent également l'amélioration des performances apportée par chaque approche de référence basée sur les RAG : GEN, Olympus et MSDA améliorent tout aussi bien le compromis entre utilité et vie privée. Cependant, notre analyse d'utilité (tableau 3.1) montre que les données assainies sont assez déformées, ce qui nuit aux traitements de signaux nécessaires pour des analyses ultérieures. MSDA intègre la distorsion des données dans sa fonction de perte, conduisant ainsi à des distorsions moins importantes. Cette fonctionnalité améliore la qualité du traitement du signal, mais n'améliore pas de manière significative le compromis entre utilité et vie privée comparativement à Olympus (figure 3.4). En sélectionnant dynamiquement le meilleur modèle d'assainissement pour chaque fenêtre de données brutes, DYSAN(o) rend l'inférence du genre proche d'une probabilité aléatoire tout en préservant une détection précise de l'activité.

Les résultats de GEN rapportés dans (MALEKZADEH, CLEGG, CAVALLARO et HADDADI, 2018) mentionnent une précision de 94% pour la reconnaissance d'activités et de 64% pour l'inférence de genre pour le jeu de données MotionSense, contre 95% et 96%, respectivement dans nos expériences. Cette différence peut s'expliquer par notre implémentation qui exploite deux réseaux de neurones, un pour chaque tâche de classification (reconnaissance d'activités et inférence de genre). Leur implémentation originale utilise un seul réseau de neurones pour les deux tâches de classification, comme expliqué dans la section 3.3.2.



(a) MotionSense



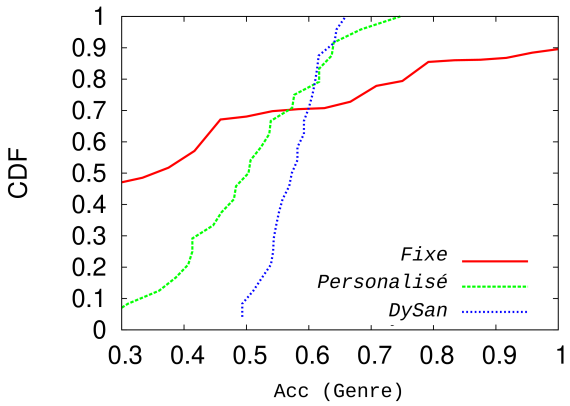
(b) MobiAct

FIGURE 3.5 – La sélection dynamique du modèle d'assainissement de DYSAN améliore significativement la reconnaissance d'activités dans le cadre d'un apprentissage par transfert (ensemble de données MobiAct).

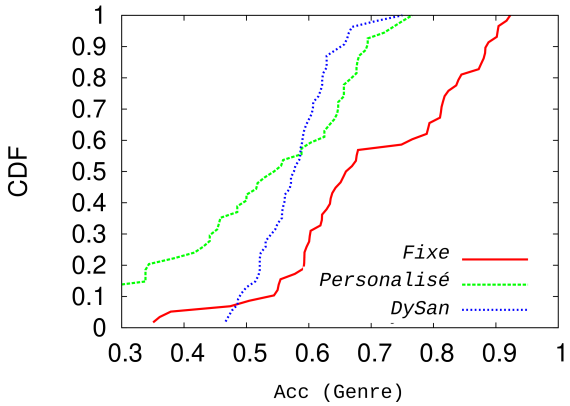
3.4.4 Sélection dynamique du modèle d'assainissement

Les figures 3.5 et 3.6 représentent pour les deux ensembles de données la distribution cumulative (*CDF*) de la reconnaissance d'activité et de l'inférence du genre, respectivement lorsqu'un modèle d'assainissement fixe, un modèle personnalisé et un modèle dynamique sont considérés. Les résultats montrent que la précision des deux tâches de classification est très hétérogène au sein de la population de personnes utilisatrices. Cette hétérogénéité élevée reflète le fait qu'un modèle statique et unique pour tous n'est pas adapté à toutes les personnes utilisatrices et encore à toutes les activités effectuées par ces derniers. Notre approche dynamique s'en trouve ainsi beaucoup plus motivée.

Plus précisément, les résultats montrent que l'adaptation dynamique du modèle d'assainissement



(a) MotionSense



(b) MobiAct

FIGURE 3.6 – En adaptant dynamiquement le modèle d'assainissement pour chaque personne utilisatrice en fonction des données entrantes, DYSAN améliore considérablement la protection contre l'inférence du genre (la distribution de la prédiction du genre est centrée autour de $0,5$, ce qui correspond à une estimation aléatoire).

améliore considérablement la reconnaissance des activités par rapport à l'utilisation d'un modèle statique en cas d'apprentissage par transfert (figure 3.5b). Pour le jeu de données MotionSense (figure 3.5a), la plupart des personnes utilisatrices obtiennent de meilleures performances avec un modèle statique fixé pour toutes les personnes utilisatrices. Ce résultat peut être expliqué par le fait que les modèles d'assainissement ont été appris avec les mêmes personnes, conduisant à un apprentissage des caractéristiques de mouvement de toutes les personnes considérées. Pour rappel, les assainisseurs sont entraînés uniquement sur l'ensemble de données MotionSense.

Pour l'inférence du genre, l'objectif de l'assainisseur est de fournir une précision autour de $0,5$, qui

correspond à des prédictions aléatoires pour tous les personnes. Cependant, les résultats décrits dans la figure 3.6 montrent qu'un modèle fixe pour toutes les personnes ne protège pas suffisamment contre l'inférence du genre. En effet, la distribution fait état d'un large éventail d'exactitude de prédiction sur lequel la déduction du genre se fait avec 80% de confiance, nous obtenons respectivement 60% et 20% des personnes utilisatrices pour les ensembles de données MobiAct et MotionSense. L'adoption d'un modèle d'assainisseur personnalisé pour chaque personne diminue la précision de la prédiction du sexe par rapport à un modèle fixe. La distribution de la précision reste toutefois importante (de $0,3$ à $0,75$ pour MotionSense et de $0,3$ à $0,8$ pour MobiAct). En adaptant dynamiquement le modèle d'assainissement en fonction des intrants, DYSAN améliore considérablement la protection contre l'inférence du genre par rapport à l'utilisation d'un modèle fixe.

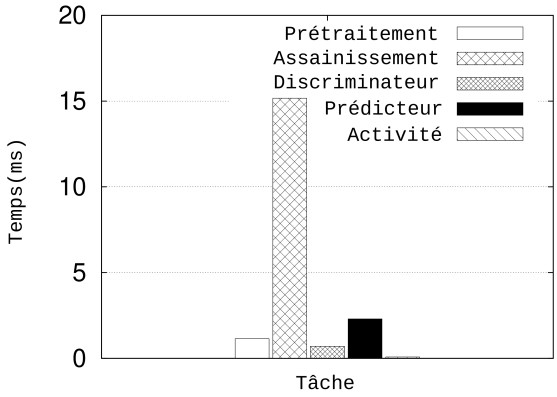


FIGURE 3.7 – La charge de calcul limitée de l'assainissement par DYSAN est compatible avec un traitement en temps réel sur téléphone intelligent.

Ces résultats montrent également la capacité de DYSAN à transférer l'apprentissage effectué sur MotionSense vers MobiAct (une reconnaissance d'activités autour de 92% en moyenne pour une prédiction du genre autour de 57%). À titre de comparaison, nous avons aussi évalué le transfert d'apprentissage de GEN en utilisant le modèle original d'assainissement entraîné sur MotionSense (et disponible publiquement) vers le jeu de données MobiAct. Dans ce cas, GEN fournit une reconnaissance d'activité et de détection de sexe d'environ 43% et 56%, respectivement. Ce résultat montre la capacité limitée de GEN à transférer l'apprentissage de MotionSense à un autre jeu de données, ce qui laisse supposer un surapprentissage du réseau de neurones sous-jacent à l'ensemble de données considéré.

Pour compléter cette analyse, nous avons aussi évalué la variation de la sélection du modèle d'assai-

nissement de DYSAN par rapport aux approches statiques ainsi que le nombre de modèles différents utilisés par DYSAN pour chaque personne (Boutet, Frindel, Gambs, Jourdan et al., 2021).

3.4.5 Performances mesurées sur les téléphones

La figure 3.7 décrit le temps (ms) passé par un téléphone Xiaomi Redmi Note 7 (équipé d'un Qualcom Snapdragon 660 et de 3 Go de mémoire, exécutant une application java avec Pytorch 1.6) sur chaque tâche associée à un seul modèle d'assainissement d'une fenêtre de données entrantes ($2,5$ secondes de données). Plus précisément, ces tâches comprennent le prétraitement du signal, l'assainissement des données brutes, l'évaluation de la vie privée et de l'utilité des données assainies, respectivement par le discriminateur et le prédicteur, et la classification de l'activité réalisée par la personne utilisatrice à partir des données brutes. À l'exception du prétraitement, qui n'est effectué qu'une seule fois pour une fenêtre de données, les autres tâches doivent être répétées pour chaque modèle d'assainissement exploré. Les résultats montrent que l'application d'un modèle d'assainissement une seule fois prend la plupart du temps, tandis que toutes les opérations nécessitent $19ms$. La prise en compte de 20 ou 36 modèles d'assainissement augmente ce temps respectivement à $366ms$ et $658ms$. Bien que ce traitement soit compatible avec un traitement en temps réel (traitement réalisé après chaque fenêtre de données), le nombre de modèles déployés sur le téléphone intelligent doit être choisi en conséquence pour limiter la surcharge. Ce nombre de modèles a également un impact sur l'espace de stockage nécessaire. En moyenne, la taille d'un seul modèle est d'environ 15 Mo. En considérant 36 modèles, on obtient 540 Mo, ce qui n'est pas une limitation par rapport aux capacités de stockage des téléphones actuels.

Nous évaluons également l'impact du nombre de modèles d'assainissement considérés. Considérer moins de modèles d'assainissement conduit à couvrir moins d'hyperparamètres et donc à limiter le compromis réalisable entre utilité et vie privée. Par conséquent, une dégradation de la précision pour la détection d'activité et l'interférence de genre est observée. Le tableau 3.3 présente les performances obtenues avec différents nombres de modèles d'assainissement disponibles pour la sélection. Les résultats montrent qu'en passant de 36 à 20 modèles d'assainissement, la précision de la reconnaissance d'activité diminue de seulement 3% et augmente de 2% l'inférence de genre.

Finalement, nous évaluons l'impact de l'exécution de DYSAN sur la consommation d'énergie du téléphone. La figure 3.8 indique la diminution de la charge de la batterie au fil du temps entre une référence où aucune opération n'est effectuée sur le téléphone intelligent, et une où avec un

	Reconnaissance d'activités (%)	Prédiction du Genre (%)
36 modèles	92	57
20 modèles	89	59
16 modèles	88	63
8 modèles	86	66

TABLEAU 3.3 – La réduction du nombre de modèles d'assainissement disponibles pour la sélection diminue la précision de la reconnaissance des activités tout en augmentant l'exactitude de prédiction du genre.

traitement en temps réel de DYSAN (c'est-à-dire qu'après chaque fenêtre de données brutes, et en explorant 36 modèles d'assainissement avant de sélectionner le meilleur). Dans les deux cas, l'écran est resté allumé pendant l'expérience. Les résultats montrent que DYSAN consomme 1% de batterie en plus après une heure, ce qui reste une consommation d'énergie raisonnable.

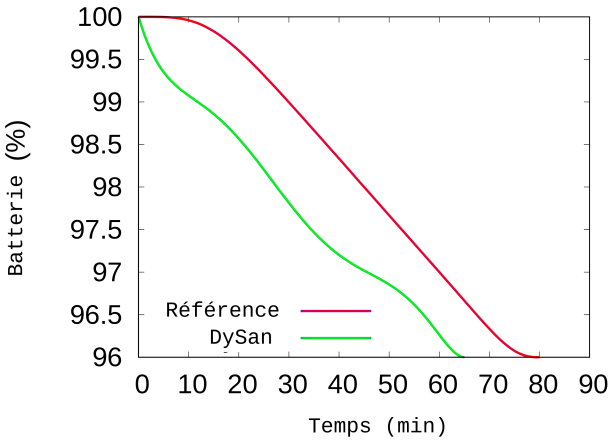


FIGURE 3.8 – L'impact de DYSAN sur la consommation d'énergie est limité (1% de batterie en moins après une heure).

3.5 DYSAN : conclusion

Nous avons présenté DYSAN, une méthode de préservation de la vie privée qui assaini les données des capteurs de mouvement afin d'empêcher l'inférence indésirable d'informations sensibles par un adversaire utilisant les données brutes. En même temps, DYSAN préserve autant que possible les informations utiles pour la reconnaissance d'activité et d'autres estimateurs de suivi de l'activité physique. Les résultats montrent que DYSAN réduit considérablement le risque d'inférence du genre, sans impacter la capacité à reconnaître l'activité ou à compter le nombre de pas. Nous montrons

également que la sélection dynamique du modèle d'assainissement de DYSAN adapte avec succès la protection à chaque personne dans le temps en fonction de l'évolution des données entrantes. Cette méthode est particulièrement efficace dans un cas d'apprentissage par transfert, contrairement aux autres méthodes qui ont un modèle d'assainissement unique pour toutes les personnes. De plus, nous avons montré que le coût supplémentaire introduit sur le téléphone intelligent pour assainir les données est compatible avec un traitement en temps réel tout en gardant une consommation énergétique raisonnable. Enfin, nous avons comparé notre approche aux approches existantes et démontré que DYSAN permet un meilleur contrôle du compromis vie privée-utilité. Nous avons également étudié la possibilité d'étendre DYSAN pour prendre en compte plusieurs attributs sensibles. Les résultats préliminaires obtenus en ajoutant plusieurs discriminateurs comptabilisés dans la fonction de perte de l'apprentissage de l'assainisseur sont encourageants, mais nous sommes limités par la petite taille des ensembles de données disponibles. En effet, rendre les modèles d'assainissement plus complexes nécessite plus de données pour capturer la spécificité de chaque cas d'utilisation.

CHAPITRE IV

FAIRMAPPING: FONCTION DE TRANSFERT DE TRAITEMENTS SPÉCIAUX

Dans le chapitre 2, nous avons introduit notre approche GANSan qui permet d'améliorer l'équité par la protection de l'attribut sensible (assainissement des données), tout en maintenant le maximum d'utilité des données ainsi assainies. Toutefois, GANSan souffre de quelques limitations dont l'introduction de modifications qui ne permettent pas de maintenir la sémantique des données (bien que ces modifications servent à protéger l'attribut sensible), et l'approche n'offre pas de garanties contre l'inférence de plusieurs groupes. Une solution simple pour protéger plusieurs attributs sensibles consisterait en l'application successive de GANSan afin de protéger chacune des informations désignées. Cependant, cette solution ne serait pas optimale étant donné que les solutions trouvées pour protéger un attribut ne correspondraient pas nécessairement aux solutions appropriées pour les autres. L'application successive de GANSan conduirait ainsi à un déplacement d'une solution locale (par rapport à un attribut) à une autre, sans toutefois atteindre la meilleure solution pour l'ensemble des attributs. De plus, le dommage introduit dans les données serait encore plus important.

Une deuxième solution possible consisterait en l'ajout d'autant de discriminateurs que d'attribut sensible, mais on peut facilement constater que cette solution passerait difficilement à l'échelle si le nombre d'attributs sensibles devient grand (temps de calcul, architecture, etc.). Cette solution nécessiterait non seulement des jeux de données de tailles conséquentes (comme évoqué dans la conclusion de DYSAN). De même, puisque chaque discriminateur serait associé à un attribut sensible, aucune garantie ne pourrait être émise concernant les sous-groupes identifiés par la combinaison des valeurs de ces attributs. Par ailleurs, plus le nombre d'attributs serait élevé, plus les sous-groupes issus des combinaisons seraient de petites tailles, rendant difficile l'apprentissage par des RAG.

Concernant le maintien de la sémantique, la plupart des approches de prétraitement des données telles que GANSan transforment les données pour protéger l'attribut sensible en déplaçant les distri-

butions des attributs vers une distribution intermédiaire dans laquelle l'appartenance aux groupes est masquée. Cette distribution intermédiaire est celle qui, selon la technique utilisée, maximise aussi l'utilité des données modifiées. Par exemple, l'approche *Disparate Impact Remover* (Feldman, Friedler, Moeller, Scheidegger et al., 2015) effectue une translation des distributions conditionnelles (par rapport à l'attribut sensible) des attributs vers une distribution médiane, le degré de translation est contrôlé par un coefficient λ. La perte de sémantique et l'introduction de modifications illégitimes proviennent du fait que cette distribution intermédiaire peut représenter une distribution qui n'existe pas dans le monde réel, et ne peut ainsi pas être utilisée à différents niveaux d'interprétations, selon la finalité de la tâche. Nous avons vu par exemple que GANSan effectue un «alignement» des données basées sur les distributions conditionnelles, qui peuvent résulter en des profils irréalistes. Par exemple, nous y avons montré comment l'attribut *relation* présenté dans le tableau 4.1 (distribution originale) et le tableau 4.2 (distribution assainie) est modifié par l'assainissement. On peut remarquer notamment que la proportion de la valeur *Mari*, qui passe de $0,0068$ à $30,43$ dans le sous-groupe de données originales *Femme*. Sémantiquement parlant, à la période de collection de ces données, l'association *Femelle-Mari* aurait pu être considérée comme un contresens, bien que nous puissions observer à partir de la distribution assainie que l'attribut sensible est protégé (la distribution des valeurs de cet attribut se rapproche de celle de l'attribut sensible se rapproche de celle de l'attribut sensible).

Tableau 4.1 – Distribution originale de l'attribut *relation* (*relationship* dans l'ensemble de données). Les valeurs numériques représentent le pourcentage de chaque valeur possible.

Attributs	**relation**					
Valeurs	**Mari**	**Hors de la famille**	**Enfant**	**Célibataire**	**Femme**	**Autre lien de parenté**
$Pr(Y = Y_x)$	$41,27$	$25,87$	$14,65$	$10,58$	$4,62$	$2,98$
$Pr(Y = Y_x \mid S = Mâle)$	$61,14$	$20,60$	$12,11$	$3,71$	$0,0033$	$2,42$
$Pr(Y = Y_x \mid S = Femelle)$	$0,0068$	$36,82$	$19,93$	$24,85$	$14,22$	$4,15$
$Pr(S = Mâle \mid Y = Y_x)$	$99,99$	$53,75$	$55,79$	$23,70$	$0,0478$	$54,78$
$Pr(S = Femelle \mid Y = Y_x)$	$0,0054$	$46,24$	$44,20$	$76,29$	$99,95$	$45,21$

Pour mieux illustrer cette problématique, considérons un ensemble de données composé de trois attributs, *origine ethnique*, *nombre d'années d'études* et *diplôme obtenu*. L'attribut sensible *origine ethnique* est considéré fortement corrélé à *nombre d'années d'études* et indépendant de l'attribut *diplôme obtenu*. Pour protéger l'attribut sensible, l'attribut *nombre d'années d'études* serait modifié

TABLEAU 4.2 – Distribution de l'attribut *relation* après assainissement.

Attributs	**relation**					
Valeurs	**Mari**	**Hors de la famille**	**Enfant**	**Célibataire**	**Femme**	**Autre lien de par.**
$Pr(Y = Y_x)$	$59,24$	$21,85$	$18,89$	$0,00$	$0,00$	$0,0044$
$Pr(Y = Y_x \mid S = Mâle)$	$61,05$	$21,23$	$17,70$	$0,00$	$0,00$	$0,0033$
$Pr(Y = Y_x \mid S = Femelle)$	$55,48$	$23,13$	$21,37$	$0,00$	$0,00$	$0,0068$
$Pr(S = Mâle \mid Y = Y_x)$	$69,56$	$65,59$	$63,24$	$0,00$	$0,00$	$0,50$
$Pr(S = Femelle \mid Y = Y_x)$	$30,43$	$34,40$	$36,75$	$0,00$	$0,00$	$0,50$
$Recall_{GB} : Pr(\hat{Y} = Y_x \mid Y = Y_x)$	$99,89$	$97,43$	$94,43$	$0,00$	$0,00$	$0,00$
$Precis_{GB} : Pr(Y = Y_x \mid \hat{Y} = Y_x)$	$99,97$	$95,09$	$97,02$	$0,00$	$0,00$	$0,00$

par les approches de prétraitement, tandis que l'attribut *diplôme obtenu* pourraient-être inchangé étant donné qu'il n'apporte pas d'information sur la décision. Les approches de prétraitement des données transformeraient ainsi les intrants en une distribution intermédiaire (située dans le même espace que les données originales) non maîtrisée, dans laquelle on pourrait retrouver des personnes de niveau *Lycée* (5 à 7 années d'études secondaires, attribut *nombre d'années d'études*) détenant un diplôme de *doctorat*. Ces discordances entre les attributs peuvent provenir du fait que certaines valeurs sont prédominantes dans l'ensemble de données, en plus de la liberté accordée au modèle dans le choix de la distribution intermédiaire (qui peut en conséquence choisir celle qui ne respecte pas la sémantique des données). En conséquence, l'utilité de l'ensemble de données est réduite puisque des différences sémantiques plus subtiles (comme dans nos exemples) peuvent conduire à des interprétations erronées des données. Enfin, la plupart des approches de prétraitement nécessitent l'attribut sensible S pour guider la transformation des données (par exemple pour connaitre la direction de la translation dans l'approche DIRM).

Notre approche Fair Mapping est introduite pour pallier ces lacunes, notamment en contraignant la distribution cible de la transformation. Fair Mapping est inspirée par l'approche AttGAN (HE, ZUO, KAN, SHAN et al., 2019) et exploite les RAG Wasserstein (WGAN) (ARJOVSKY, CHINTALA et BOTTOU, 2017) pour réaliser le *transport optimal* des intrants vers la distribution cible choisie. À cela, nous rajoutons des contraintes de vie privée pour prévenir l'inférence de l'attribut sensible. Ce transport de distribution vers une distribution cible connue et existante dans le monde réel permet de maintenir l'aspect réaliste des données, et offre à notre outil la possibilité d'être utilisé pour la détection de discrimination. De même, puisque notre approche s'appuie sur la théorie du transport optimal, elle bénéficie de ses propriétés, notamment l'introduction du déplacement minimal

pour transformer une distribution en une autre. Enfin, notre approche ne requiert aucunement la connaissance de l'attribut sensible dans sa phase de test et de déploiement, ce qui permet ainsi aux populations sensibles de ne pas révéler cette information une fois notre système déployé. En effet, l'attribut sensible dans cette approche est utilisé uniquement durant la phase d'entraînement afin de distinguer le groupe privilégié (qui sera considéré comme groupe cible) des autres groupes de l'ensemble d'entraînement. À partir des propriétés de notre approche, la distinction entre les groupes n'est pas nécessaire durant la phase de test et de déploiement.

Une fois le modèle entraîné, FairMapping pourrait être utilisé dans deux cas d'utilisation distincts.

— L'approche pourrait être utilisée par un conservateur de données qui transformerait les informations à sa disposition afin d'assurer un traitement similaire à tous, réduisant ainsi les risques de discrimination.

— Dans une autre situation, l'approche pourrait être utilisée localement par des personnes qui souhaitent collaborer par le partage de leurs informations, mais veulent s'assurer que leurs attributs sensibles ne puisse être inférés tout en bénéficiant des mêmes avantages que les personnes du groupe cible. Ici, le code de *FairMapping* pourrait être fourni par une entité indépendante.

Notre article FairMapping a été soumis dans une édition spéciale sur la thématique *Safe and Fair Machine Learning* du journal *Machine Learning* et est en cours de révision. Je suis l'auteur principal de cet article et j'y ait notamment contribué par la conception de la méthode et du modèle, la réalisation des expérimentations et la rédaction de l'article.

4.1 Fair mapping : cadriciel

Dans cette section, nous présentons notre approche de prétraitement, appelée *Fair Mapping*, dont l'objectif est d'apprendre à transformer toute distribution de l'ensemble de données en une distribution cible choisie. Pour être plus précis, notre objectif est de protéger le ou les attributs sensibles en apprenant une fonction de transformation des données de n'importe quel groupe de l'ensemble de données (groupes protégés ou groupe privilégiée) en des versions appartenant à la distribution cible (notamment la distribution privilégiée), mais à partir desquelles il serait difficile d'inférer l'attribut sensible. Nous envisageons deux cas d'utilisation potentiels de notre approche :

— La fonction de transformation vers le groupe cible peut être utilisée par une entité de gestion centralisée des données pour étendre le traitement spécifique d'un groupe de l'ensemble de

données aux autres groupes de cet ensemble, réduisant ainsi le risque de traitement différentiel entre ces groupes.

— La personne utilisatrice pourrait appliquer localement la fonction de transformation pour assainir son profil avant sa publication, ceci afin de s'assurer que les attributs sensibles sont protégés contre les attaques par inférence, tout en bénéficiant du même traitement (privilèges, voire désavantages) que celui réservé aux membres du groupe cible. Dans cette situation, la fonction pourrait être fournie par une entité indépendante aux personnes préoccupées par l'utilisation abusive de leurs données ou craignant d'être victimes de discrimination en raison de leur appartenance de groupe.

L'objectif de notre approche diffère de celui des techniques de prétraitement pour l'amélioration de l'équité (partie 1.6.2) sur plusieurs points. Premièrement, notre approche transforme tous les profils de l'ensemble de données vers la distribution privilégiée choisie, garantissant que tous les groupes bénéficient du même traitement, contrairement aux techniques de la littérature qui transforment les données vers une distribution intermédiaire proche de la médiane. En conséquence, notre transformation préserve l'aspect réaliste de l'ensemble de données puisque la distribution privilégiée est une distribution réelle (la distribution intermédiaire qui ne représente ni les groupes protégés ni la distribution privilégiée). De plus, la transformation de tous les profils vers une distribution connue permet de valider le processus de transformation, notamment en s'assurant que les données provenant de la distribution privilégiée ne soient pas affectées par la transformation. Enfin, les données transformées peuvent également être utilisées pour la détection de la discrimination, en observant par exemple si la décision change selon que l'on utilise le profil du groupe protégé ou sa version obtenue dans le groupe privilégié, comme cela est réalisé dans FlipTest (Black, Yeom et Fredrikson, 2020).

4.1.1 Vue d'ensemble de Fair Mapping

Comme mentionné dans l'introduction, notre objectif est d'apprendre une *fonction de transformation* qui peut être appliquée à différents groupes de l'ensemble de données de sorte que la quantité de modifications introduites soit minimale, que tous les profils de données transformés appartiennent à la distribution privilégiée (donc, partagent les mêmes caractéristiques ou privilèges), et que l'attribut sensible ne puisse pas être inféré. Ces objectifs peuvent se traduire respectivement en propriétés mesurables d'*identité*, de *transformation* et de *protection* :

1. *Propriété d'identité.* Idéalement, la transformation ne devrait pas modifier les profils des per-

sonnes qui appartiennent déjà à la distribution privilégiée (R_{priv}). En fait, comme l'objectif est de transformer tous profils de l'ensemble de données vers la distribution privilégiée en trouvant leur version *privilégiée* correspondante, la transformation qui produit le moins de modification pour une donnée r_i provenant de la distribution privilégiée, est cette donnée r_i elle-même. À partir de cette observation, la propriété d'identité peut être utilisée pour mesurer la qualité de transformation de tous les groupes, pour deux raisons principales. Tout d'abord, toute transformation optimale qui produit le moins de modifications sur l'ensemble des données produira également le moins de modifications dans la transformation du groupe privilégié. Ceci est dû au fait que la minimisation de la quantité de modifications sur l'ensemble des données exige que tous les groupes soient transformés avec la moindre quantité de modifications possible ; Ensuite, bien que la réciproque de l'affirmation précédente (*une transformation avec le moins de modifications sur le groupe privilégié fournira également le moins de modifications sur la transformation des groupes protégés*) n'est pas nécessairement vrai, la propriété d'identité peut servir de substitut pour évaluer la qualité de notre transformation. En effet, comme la version privilégiée des données protégées qui nécessitent le moins de modifications n'est pas connue à priori, on ne peut mesurer la qualité de la transformation de ce groupe pour s'assurer qu'il s'agit bel et bien de la meilleure parmi toutes les transformations possibles.

2. *Propriété de transformation.* Du point de vue de la distribution des données originales, un profil transformé devrait être prédit comme faisant partie du groupe privilégié. Cette propriété indique que tout classifieur entraîné sur les données d'origine pour inférer l'attribut sensible devrait prédire que tout profil transformé appartient à la distribution du groupe privilégié. Comme l'identité garantit déjà que le groupe privilégié n'est pas modifié, la portée de cette propriété pourrait être limitée aux données des groupes protégés R_{prot}.Grâce à cette propriété, les modèles déjà entraînés sur les données originales pour une tâche spécifique n'ont pas besoin d'être réentraînés lorsqu'ils utilisent les données transformées. En effet, l'ensemble de données sur lequel ces modèles sont entraînés contient déjà la distribution privilégiée. Ainsi, en transformant tous les profils de données vers la distribution privilégiée, nous transformons les données sur une distribution déjà connue par les modèles entraînés, aucune modification supplémentaire n'est donc nécessaire du côté de ces modèles.

3. *Protection de l'attribut sensible.* En supposant qu'il existe plus d'un groupe protégé (par exemple, deux groupes protégés), la transformation de ces groupes vers la distribution privi-

légiée devrait produire des résultats dans lesquels les groupes protégés sont indiscernables. En fait, la transformation des distributions protégées vers la distribution privilégiée ne suffit pas à garantir la protection de l'attribut sensible. À titre d'illustration, supposons que la distribution privilégiée corresponde à une distribution gaussienne bimodale. Une transformation qui fait correspondre exclusivement le premier groupe protégé au premier mode de distribution privilégiée (gaussiènne) tout en faisant correspondre l'autre groupe protégé au deuxième mode de distribution privilégiée satisferait l'objectif de transformation (puisque chaque mode fait partie de la distribution bimodale), mais un classifieur serait toujours capable de construire une frontière de décision entre les deux modes (*cf.* figure 4.1).

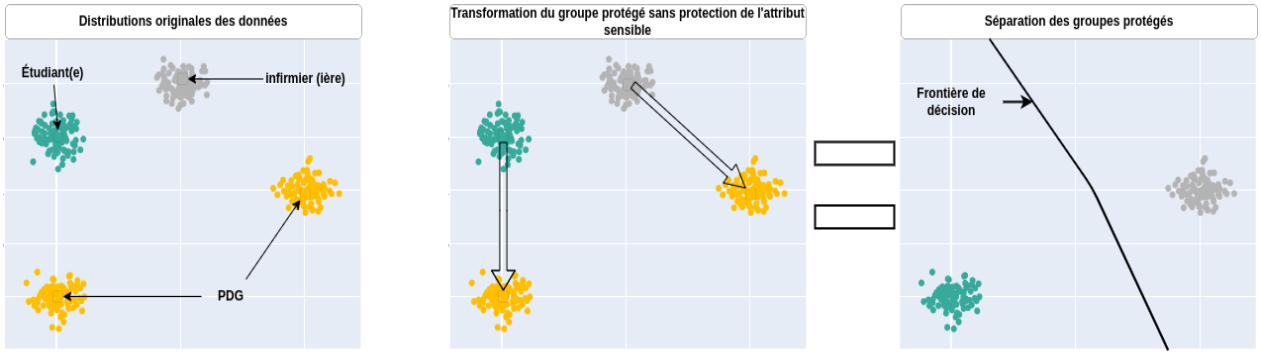


FIGURE 4.1 – Transformation des groupes protégés (*infirmiers(ères)* et *étudiants(es)*) sur la distribution privilégiée *pdg*. Les données transformées appartiennent à la distribution cible, mais l'attribut sensible n'est pas protégé.

De même, un mécanisme de transformation qui fait correspondre tous les groupes protégés au premier mode de la distribution privilégiée garantirait que tous les groupes protégés sont indiscernables les uns des autres, mais ceux-ci sont toujours distinguables du second mode de la distribution privilégiée, donc distinguables de la distribution privilégiée. Un classifieur pourrait également construire une frontière de décision entre la distribution protégée transformée et la distribution privilégiée.

La propriété de protection doit supprimer toute dissemblance entre la distribution privilégiée et les distributions protégées transformées. Dans l'exemple précédent, la protection garantit que chaque distribution du groupe protégé soit également une gaussienne bimodale avec les mêmes statistiques que la distribution privilégiée (*cf.* figure 4.2).

Notre approche FM s'appuie sur les WGAN afin de satisfaire nos objectifs. Les éléments constitutifs

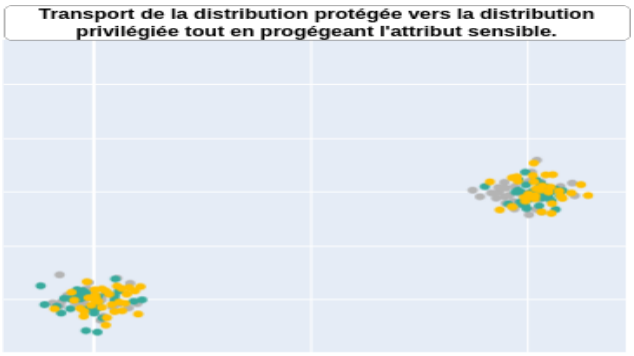


FIGURE 4.2 – Transformation des groupes protégés (*infirmiers(ères)* et *étudiants(es)*) sur la distribution privilégiée *pdg*. L'attribut sensible est protégé.

de notre approche sont le modèle de transformation G_w, le discriminateur D, la critique D_{std} et le classifieur C. L'utilisation du WGAN (instancié dans notre cadre avec le jeu à deux joueurs entre G_w et D_{std}) garantit que seules les modifications nécessaires à la transformation sont introduites lors de la transformation des données protégées, comme montré dans la partie 1.7.

Considérons un ensemble de données composé de trois classes sensibles, s_0 = étudiant(e), $s_1 = pdg$ et s_2 = infirmier(ière), avec des personnes de chaque domaine effectuant une demande de prêt. L'objectif du FM sera pour le modèle G_w de transformer les profils de données de toutes les classes vers le domaine *pdg* (R_{pdg}), de sorte que les personnes des classes *infirmier(ière)* et le *étudiant(e)* bénéficient du même traitement (figure 4.3) que celles de la classe *pdg*. De plus, G_w doit se comporter comme la fonction d'identité pour tous profil du domaine *pdg*, et assurer que les transformations de tous les groupes sont indiscernables les unes des autres. L'aperçu de haut niveau de notre approche FM consiste :

1. Pour chaque profil r_i^t dans tout domaine $t \in$ {pdg, étudiant(e), infirmier(ère)}, générer la version transformée $\bar{r}_i^t = G_w(r_i^t)$.
2. S'assurer que $\bar{r}_i^t = r_i^t$ si t est égal à *pdg* (identité).
3. S'assurer, si t est différent de *pdg*, que $\bar{r}_i^t$ appartient au domaine *pdg* en utilisant le classifieur C et la critique D_{std}, et que la transformation a été effectuée avec le moins de modifications possible.
4. S'assurer, avec l'aide du discriminateur D, que toutes les sorties du modèle G_w sont indiscernables les unes des autres.

5. Mettre à jour les modèles en fonction des observations précédentes.

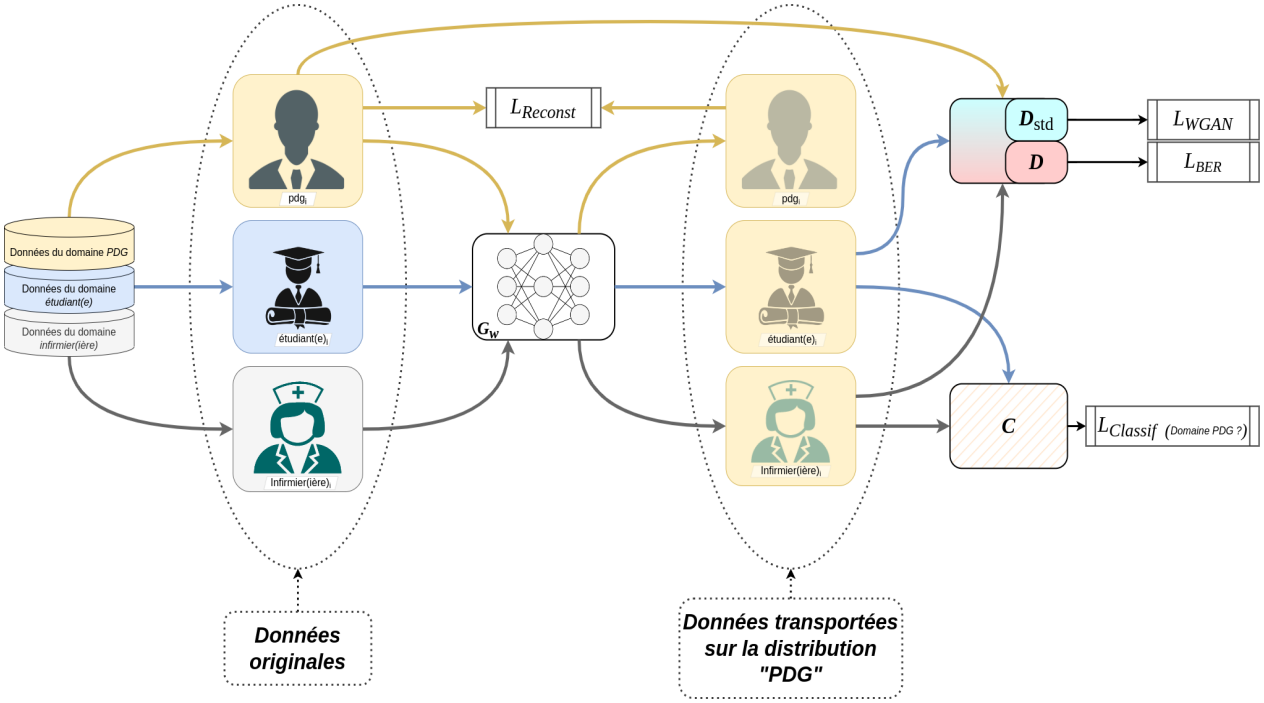


FIGURE 4.3 – Approche FairMapping (FM). L'objectif est de transformer tous les autres domaines de données considérés comme protégés (domaine de données de l'étudiant(e) et domaine de données de l'infirmièr(ère)) sur le domaine de données *pdg* (représenté par le fond jaune), de sorte que tous les profils de données partagent les mêmes avantages (par exemple dans le cas d'une demande de prêt) que le *pdg*, de sorte que l'étudiant(e) transformé(e) et l'infirmier(ère) transformé(e) soient considérés comme faisant partie du groupe privilégié du point de vue du classifieur C, et de sorte que le discriminateur D soit incapable de distinguer le *pdg*, de *infirmier(ère)* et de *étudiant(e)*.

4.1.2 Procédure d'entraînement

Nous décrivons ci-après nos modèles et leur procédure d'apprentissage respective. Nous présentons tout d'abord l'entraînement de C qui se fait avant tout autre modèle, suivi de l'entraînement du discriminateur D, de la critique D_{std}, et enfin l'entraînement de G_w. Les trois derniers modèles sont entraînés alternativement, tout comme dans l'entraînement des RAGs.

Entraînement du classifieur C. La procédure d'apprentissage de FM commence par l'apprentissage des classifieurs C pour prédire l'attribut sensible. Pour chaque profil r_i^t avec une valeur d'attribut

sensible de s_t, C fournit les probabilités c_i^j, qui correspondent aux probabilités que r_i^t appartienne à chaque groupe sensible y_j, $C(r_i^t) = \langle c_i^1, \ldots, c_i^t, \ldots, c_i^k \rangle$; $c_i^j = C(r_i^t)^j$ et $\sum_{j=1}^{k} c_i^j = 1 \, \forall \, i$, l'objectif étant de maximiser c_i^t (la probabilité d'appartenance au groupe y_t pour le profil r_i^t). C est entraîné en utilisant les données originales jusqu'à convergence, son objectif est de maximiser la précision de la prédiction de l'attribut sensible :

$$L_{Classif} = L_{Acc} = 1 - \frac{\sum_{r_i^j \in R} C(r_i^j)^j}{\#(R)}. \tag{4.1}$$

où $\#(R)$ représente la cardinalité de l'ensemble de données, c'est-à-dire $N = \#(R)$. Puisque C est entraîné sur les données originales qui sont indépendantes du processus de transformation, la procédure d'entraînement de C est indépendante de celle des autres modèles. Ainsi, C peut donc être entraîné jusqu'à convergence avant son utilisation dans le processus de transformation.

Apprentissage du modèle D_{std} et du discriminateur D

Le modèle D_{std} correspond à celui utilisé pour l'apprentissage de WGAN. Il prend en entrée les données originales de la distribution privilégiée, les données protégées transformées, et doit donner une valeur correspondante à une mesure de *distance* entre les deux types de données (équation 4.2). L'objectif, tout comme dans le cas de WGAN, est de trouver le modèle qui maximiserait la *distance* calculée $L_{D_{std}}$ entre le groupe privilégié et les groupes protégés transformés. La recherche d'un tel modèle permet d'introduire le minimum de modifications nécessaires à la transformation des données protégées.

$$\begin{aligned} L_{WGAN} = L_{D_{std}} &= D_{std}(R_{priv}) - D_{std}(G_w(R_{prot})) \\ &= \frac{\sum_{r_i \in R_{priv}} D_{std}(r_i)}{\#(R_{priv})} - \frac{\sum_{r_j \in R_{prot}} D_{std}(G_w(r_j))}{\#(R_{prot})}. \end{aligned} \tag{4.2}$$

Dans l'équation 4.2 $\#(R_{priv})$ et $\#(R_{prot})$ représentent respectivement la taille du groupe privilégié et celle du groupe protégé. Le discriminateur D est similaire au classifieur C dans la prédiction de l'attribut sensible. Toutefois, D prédit l'attribut sensible sur la base des données protégées transformées (obtenues du modèle G_w) et des données privilégiées originales. Ainsi, D est entraîné pour maximiser l'identification de chaque groupe protégé transformé ainsi que du groupe privilégié original. La fonction de perte pour l'apprentissage de D est similaire à celle de C, à savoir la minimisation de la perte d'exactitude de prédiction (L_{Acc}, équation 4.3). D peut également être entraîné avec l'objectif de minimiser le BER (équation 4.4), qui correspond à la minimisation de l'erreur de prédiction

dans chaque groupe. Nous désignerons la fonction de perte du discriminateur par L_D : $L_D = L_{Acc_D}$ ou $L_D = L_{BER_D}$.

$$L_{Acc_D} = 1 - \frac{\sum_{r_i^1 \in R_{priv}} D(r_i^1)^1 + \sum_{r_i^j \in R_{prot}} D(G_w(r_i^j))^j}{\#(R)}. \tag{4.3}$$

.

$$L_{BER_D} = 1 - \frac{1}{k}\left(\frac{\sum_{r_i^1 \in R_{priv}} D(r_i^1)^1}{\#(R_{priv})} + \sum_{j=2}^{k} \frac{\sum_{r_i^j \in R_{prot}|S=s_j} D(G_w(r_i^j))^j}{\#(S=s_j)}\right). \tag{4.4}$$

En cas de groupes multiples, nous calculons également l'information mutuelle (MI) entre la prédiction de l'attribut sensible réalisée avec D, $\bar{S}$, et sa valeur réelle S. L'information mutuelle est introduite comme une régularisation de D pour améliorer la prédiction de groupes multiples et doit être maximisée.On obtient donc la fonction de perte du discriminateur $L_D = L_{Acc} - \lambda_{D_{mi}} MI$.

Le discriminateur D et le critique D_{std} partagent un ensemble de paramètres, ce partage semble conduire à une relation bénéfique entre les deux modèlesd'après nos premières expérimentations). En conséquence, l'apprentissage des deux modèles D et D_{std} est régit par la fonction de perte : $L_D = L_D + \lambda_{D_{stdgan}} * L_{D_{std}}$.

Entraînement du modèle G_w. *Identité.* Pour assurer la non modification des membres du groupe privilégié R_{priv}, nous nous appuyons sur la fonction de reconstruction L_{Recons}. Cette fonction, instanciée avec la norme $L1-$ (équation 4.5), contraint le modèle G_w à apprendre la distribution des données privilégiées. Ainsi, G_w est capable d'apprendre l'espace des caractéristiques de cette distribution. En fait, nous avons observé dans nos premières expériences que sans la contrainte de reconstruction, le modèle G_w est fortement sujet au phénomène de *mode collapse*, où le modèle génératif n'apprend qu'un seul ou très peu de modes de la distributions de données. Le modèle, par exemple, produit le même résultat identique quelle que soit l'intrant.

$$L_{Recons} = L_1(R_{priv}, G_w(R_{priv})) = \frac{\sum_{r_i \in R_{priv}} \mid r_i - G_w(r_i) \mid}{\#(R_{priv})}. \tag{4.5}$$

La reconstruction renforce donc la diversité des transformations, tout en améliorant la capacité de G_w à servir de fonction d'identité pour les membres du groupe privilégié.

Transformation. Rappelons que le groupe privilégié est associé à la valeur d'attribut sensible s_1. Le mécanisme G_w doit transformer les données pour qu'elles appartiennent au groupe privilégié.

Ainsi, C est utilisé pour assurer cette propriété en donnant la probabilité qu'un profil r_i du groupe protégé appartienne au groupe cible après être transformé, $C(G_w(r_i))^1$. Par conséquent, G_w vise à maximiser la probabilité $C(G_w(r_i))^1$. La sortie du classifieur C est intégrée dans l'apprentissage de G_w par le biais de la fonction de perte L_C (équation 4.6) :

$$L_C = \frac{\sum_{r_i \in R_{prot}} C(G_w(r_i))^1}{\#(R_{prot})}. \tag{4.6}$$

La transformation est encore améliorée avec l'utilisation du transport optimal, comme expliqué dans le cadre des WGAN, dans la section 1.7. Le WGAN garantit en outre que seules les modifications nécessaires à la transformation sont introduites pendant celle-ci. Ainsi, G_w est aussi entraîné de manière similaire aux WGAN, qui correspond à la minimisation de l'équation 4.2. Le premier terme de l'équation 4.2, $D_{std}(R_{priv})$, ne fait pas intervenir G_w, ses dérivées par rapport aux paramètres G_w sont égales à zéro. Nous ne pouvons donc que minimiser le terme $-D_{std}(G_w(R_{prot}))$ par rapport aux paramètres du modèle G_w. Nous pouvons écrire ainsi l'équation 4.7

$$L_{Gan} = -D_{std}(G_w(R_{prot})) = -\frac{\sum_{r_j \in R_{prot}} D_{std}(G_w(r_j))}{\#(R_{prot})}. \tag{4.7}$$

L'objectif est d'avoir une distance de 0 mesurée par D_{std} entre les données privilégiées R_{priv} et les protégées transformées $G_w(R_{prot})$.

Protection. La protection consiste en l'impossibilité d'inférer l'attribut sensible, à partir des données privilégiées (R_{priv}) et des données protégées transformées ($G_w(R_{prot})$). Cette propriété est introduite dans l'optimisation de G_w par la perte L_S, qui peut être instanciée avec le BER à maximiser jusqu'au niveau de protection idéal (minimisation de l'équation 4.8), ou l'exactitude de prédiction à minimiser (minimisation de l'équation 4.10).

$$\begin{aligned} L_S = &\frac{k-1}{k} - \\ &\frac{1}{k}\left(1 - \frac{\sum_{r_i^1 \in R_{priv}} D(r_i^1)^1}{\#(R_{priv})} + \sum_{j=2}^{k}\left[1 - \frac{\sum_{r_i^j \in R_{prot}|S=s_j} D(G_w(r_i^j))^j}{\#(S=s_j)}\right]\right). \end{aligned} \tag{4.8}$$

Après simplification, l'équation L_S s'écrit sous la forme suivante :

$$\begin{aligned} L_S = &\frac{-1}{k} + \\ &\frac{1}{k}\left(\frac{\sum_{r_i^1 \in R_{priv}} D(r_i^1)^1}{\#(R_{priv})} + \sum_{j=2}^{k}\frac{\sum_{r_i^j \in R_{prot}|S=s_j} D(G_w(r_i^j))^j}{\#(S=s_j)}\right). \end{aligned} \tag{4.9}$$

L_S instanciée avec la précision est présenté dans l'équation suivante :

$$L_S = \frac{\sum_{r_i^j \in \{R_{priv}, G_w(R_{prot})\}} D(r_i^j)^j}{\#(R_{priv})}. \tag{4.10}$$

Dans le cas d'un attribut binaire, minimiser l'exactitude de prédiction ne protège pas l'attribut sensible, car l'obtention d'une exactitude de prédiction de zéro signifie simplement que les prédictions du modèle sont inversées par rapport aux valeurs réelles. Le BER pourrait être plus appropriée comme fonction de perte dans ce cas. Lorsqu'on a affaire à plusieurs groupes protégés, on introduit également dans l'étape de *protection* l'information mutuelle, qui doit être minimisée. L'information mutuelle est particulièrement utile car il peut exister une situation dans laquelle l'exactitude de prédiction est minimale (ou le BER est maximal) pour la protection, mais la quantité d'information concernant l'attribut sensible, préservées dans les données, est encore importante. Par exemple, le discriminateur pourrait prédire que chaque membre du groupe *pdg* appartient au groupe *étudiant(e)*, chaque *étudiant(e)* appartient au groupe *infirmier(ère)* et chaque *infirmier(ère)* est un *pdg*. Dans ce cas, l'exactitude de prédiction est égale à zéro, mais les informations sur l'appartenance au groupe ne sont pas aléatoires. Il est donc nécessaire de minimiser l'information mutuelle entre les groupes sensibles prédits et les groupes réels. La protection est donc contrôlée avec la perte :

$$L_S + \lambda_{G_{mi}} MI \tag{4.11}$$

L'information mutuelle permet également d'utiliser l'exactitude de prédiction comme fonction de perte dans le cas d'un seul attribut sensible binaire, car il ne serait pas possible d'obtenir une précision de 0 avec une information mutuelle égale à 0.

Optimisation. Chaque composante de l'optimisation de G_w étant définie, nous pouvons introduire la fonction de perte globale que le modèle de transformation G_w doit minimiser :

$$L_{G_w} = \lambda_{rec} L_{Recons} - \lambda_c L_C + \lambda_{gan} L_{Gan} + \lambda_d L_S \tag{4.12}$$

Dans cette équation, on peut observer que chaque terme de perte est pondéré par un coefficient λ. Ces coefficients permettent d'attribuer l'importance relative de chaque terme, et sont utilisés pour mieux affiner notre approche, en fonction de plusieurs facteurs tels que l'ensemble de données, le nombre de groupes, etc.

Le code de notre approche est disponible à l'adresse suivante : `https://gitlab.privsec.ca/Rosin/fairmapping`

4.2 FM : cadre expérimental

Dans cette section, nous décrivons le cadre expérimental utilisé pour évaluer notre approche. Tout d'abord, nous présentons les jeux de données utilisés, puis les approches de l'état de l'art auxquelles nous nous comparons. Ensuite, nous discuterons des métriques d'évaluation et de l'ensemble des classifieurs externes que nous utilisons durant nos expérimentations. Enfin, nous passerons en revue les différents cas d'utilisation dans lesquels notre approche peut être déployée.

Nous évaluons l'approche sur trois ensembles de données issus de la littérature : *Adult Census Income*, *German Credit* et *Lipton*. Ces ensembles étant présentés en details dans le chapitre 1, en partie 1.5, nous en ferons un bref rappel ci-après.

Adult Census Income est composé de 45222 personnes et de 14 attributs qui décrivent le statut social et économique de chaque personne. La tâche de l'ensemble de données est la prédiction du niveau de revenu, et l'attribut sensible est l'attribut binaire *sexe*, contenant soit *Mâle* soit *Femelle*. Pour le cas multivalué, nous utiliserons aussi l'attribut *origine ethnique* avec les valeurs *Blanc* et $Non - Blanc$. Nous obtiendrons ainsi un attribut unique qui contiendra la combinaison de chaque attribut binaire (*Blanc-Mâle*, $Non - Blanc - Femelle$, etc.).

L'ensemble de données *Lipton* est un ensemble synthétique formé par 2000 profils décrits par les attributs *genre*, *longueur des cheveux*, *experience de travail* et l'attribut de décision indiquant si la personne candidate devrait être recrutée. L'attribut sensible dans cet ensemble est le genre.

German Credit est composé de 1000 personnes décrites par 21 attributs bancaires. L'attribut sensible est l'*âge*, transformé en attribut binaire à la suite d'un seuillage à la valeur de 25, maximiserait la discrimination. La décision dans cet ensemble est la qualité des personnes en tant que client (bon ou mauvais payeur).

Les performances de FM sont évaluées en deux étapes : l'étape de comparaison avec l'état de l'art et l'évaluation de l'amélioration des autres métriques d'équité. Dans ces expériences, nous avons considéré que l'attribut de décision n'était pas modifié par notre approche (nous avons utilisé la décision originale de l'ensemble de données). Par la suite, nous présentons les résultats de la comparaison avec l'état de l'art.

4.2.1 Comparaison avec l'état-de-l'art

Dans l'étape de comparaison, nous évaluons les performances obtenues avec notre approche et les comparons à celles obtenues avec plusieurs approches de l'état de l'art, à savoir WGAN (Arjovsky, Chintala et Bottou, 2017), AttGAN (He, Zuo, Kan, Shan et al., 2019), GANSan (Aïvodji, Bidet, Gambs, Ngueveu et al., 2021) et DIRM (Feldman, Friedler, Moeller, Scheidegger et al., 2015). Nous avons réimplémenté ces approches (en utilisant les mêmes procédures décrites dans les articles originaux) afin de les appliquer dans notre contexte, à l'exception du DIRM que nous avons repris du cadre AIF360 (Bellamy, Dey, Hind, Hoffman et al., 2018). Ces approches ont été décrites dans le chapitre 1.

Pour chacune de ces approches, nous calculons les métriques *fidélité*, $classification$, *BER* et *exactitude de prédiction*. Chacune de ces métriques fait écho aux différents objectifs de FM décrits dans la partie 4.1.

— La *fidélité* (Fid) représente la proximité des données modifiées avec leur version originale, une fidélité parfaite (les données sont identiques) ayant une valeur de 1 :

$$Fid = \frac{1}{\#(X)}(X - G_w(X))^2 = \frac{\sum_{r_i \in X}(r_i - G_w(r_i))^2}{\#(X)} \tag{4.13}$$

Tout au long de notre analyse, nous avons calculé la Fid de trois manières : au niveau de l'ensemble des données ($X = R$, $Fid = Fid_{all}$), uniquement au niveau des données du groupe privilégié ($X = R_{priv}$, Fid_{priv}), ou uniquement avec les données protégées ($X = R_{prot}$, Fid_{prot}).

— La $classification$ (Pc) mesure la proportion des données transformées du groupe protégé qui appartiennent au groupe privilégié, du point de vue de l'ensemble de données original. La proportion est donnée par un classifieur entraîné sur la version originale de l'ensemble de données. Dans nos expériences, nous mesurons les proportions de personnes de l'ensemble de données transformé qui ont été prédites par des classifieurs externes (décrits dans les paragraphes suivants) comme faisant partie du groupe privilégié. La $classification$ est définie de la manière suivante :

$$Pc = P(f(G_w(X)) = s_1) = \frac{\sum_{r_i \in X} f(G_w(r_i)) == s_1}{\#(X)} \tag{4.14}$$

Tout comme pour Fid, nous considérons que la classification est calculée soit au niveau de l'ensemble des données ($X = R$, $Pc = Pc_{all}$), soit uniquement au niveau du groupe protégé

$(X = R_{prot},\ Pc = Pc_{prot})$.

— Le *taux d'erreur équilibré* (BER) et le *exactitude de prédiction* ($SAcc$) sont des mesures de performance du classifieur externes que nous appliquons à la prédiction de l'attribut sensible. Comme FM à la fois protège l'attribut sensible en utilisant les données privilégiées originales (rappelons que le discriminateur D est entraîné en utilisant les données protégées transformées et les données privilégiées originales) et reconstruit les données du groupe privilégié, nous pouvons mesurer la protection de l'attribut sensible par rapport aux données privilégiées originales ou à leur version reconstruite. Dans le premier cas, nous ajouterons l'indice og_{prv} aux métriques (par exemple, $BER_{og_{prv}}$), et le second cas sera identifié par rc_{prv} ($BER_{rc_{prv}}$).

Optimisation and validation. L'optimisation de nos hyperparamètres a été réalisée à l'aide de l'outil ray-tune (Liaw, Liang, Nishihara, Moritz et al., 2018), avec optuna (Akiba, Sano, Yanase, Ohta et al., 2019) comme algorithme sous-jacent. Chaque ensemble de données est divisé en un ensemble d'entraînement et un ensemble de validation, et pour chaque approche, nous avons testé 100 combinaisons d'hyperparamètres. Nos expériences ont été réalisées sur un maximum de 6 CPUs et 2 GPUs disposant de 4 Go de mémoire chacun.

Nous entraînons chaque approche pour optimiser leur fonction objective respective :

— FM maximise Fid_{priv} pour l'identité, Pc_{prot} pour la transformation et $BER_{og_{prv}}$ pour la protection (nous pouvons également minimiser $SAcc_{og_{prv}}$ à la place de $BER_{og_{prv}}$).

— WGAN ne transforme les données que vers le groupe privilégié. Ainsi, l'approche est entraînée uniquement pour maximiser la métrique de transformation mesurée sur toutes les éléments de l'ensemble de données Pc_{all}.

— GANSan protège l'attribut sensible en trouvant la quantité minimale de perturbation à introduire dans tous les profils de l'ensemble de données. Ainsi, GANSan maximise la fidélité sur tout l'ensemble de données Fid_{all} et la protection mesurée avec le groupe privilégié reconstruit et les données protégées modifiées $BER_{rc_{prv}}$. Pour les attributs sensibles multivalués, nous exploitons le BER pour étendre l'approche $GANSan$.

— AttGAN transporte les données privilégiées sur la distribution du groupe protégé, et les données protégées sur la distribution privilégiée, tout en s'assurant que les données mappées sur leur distribution originale ne soient pas modifiées. Ainsi, il maximise Fid_{all}, et maximise également la classification des profils de données transformés dans différents groupes sur

la base de permutations aléatoires de l'attribut sensible. Nous reporterons uniquement la métrique de classification calculée par rapport au groupe cible Pc_{all}, au lieu de la permutation aléatoire.

— DIRM ne cherche à optimiser aucune métrique particulière et la protection de l'attribut sensible est contrôlée par un seul paramètre λ. Ainsi, nous avons décidé de maximiser la métrique Fid_{all} et la protection basée sur les données reconstruites $BER_{rc_{prv}}$, puisque l'approche modifie tous les groupes dans l'ensemble de données. L'approche est spécifiquement conçue pour traiter un seul attribut sensible binaire. Elle ne peut donc pas être étendue au cas multivalué.

Pour réaliser une comparaison équitable entre toutes les approches, nous nous appuyons sur un ensemble de classifieurs externes pour calculer les métriques à optimiser telles que la $classification$ et la $protection$. En effet, la plupart des approches auxquelles nous nous comparons possède leur propre modèle de discriminateur adapté à son objectif spécifique. De plus, certaines approches, comme l'approche DIRM, n'incluent pas de discriminateur pour mesurer la protection de l'attribut sensible. Nous utilisons ainsi les classifieurs externes pour avoir une base de comparaison similaire à tous.

Une fois que chaque approche a été optimisée pour son ensemble respectif de métriques, nous construisons le front de Pareto qui présente les différents compromis que nous pouvons obtenir entre les différentes métriques (par exemple, le coût en Fid_{priv} pour fournir une meilleure protection $BER_{og_{prv}}$ et Pc_{prot} pour FM, le coût sur Fid_{all} pour atteindre un $BER_{rc_{prv}}$ plus élevé pour $GANSan$). Comme chaque approche a un ensemble de métriques différentes, nous fournissons une comparaison équitable en construisant un deuxième front de Pareto au-dessus du premier. Le premier front de Pareto est spécifique à l'ensemble de métriques de l'approche, tandis que le second front de Pareto affiche les compromis obtenus dans le cadre d'une autre approche. On évalue ainsi chaque approche dans son propre contexte, ainsi que dans celui des autres approches. Ainsi, la meilleure de toutes les approches pourrait surpasser les autres dans tous les contextes étudiés, ce qui se traduirait par l'obtention de la courbe de Pareto la plus haute pour toutes les métriques maximisées.

Par exemple, pour comparer FM et GANSan, nous optimisons chaque approche avec leur ensemble respectif de métriques (décrit précédemment), et construisons leur front de Pareto respectif (Fid_{priv}, $BER_{og_{prv}}$ et Pc_{prot} pour FM et Fid_{all}, $BER_{rc_{prv}}$ pour GANSan). Cela représente des situations où chaque approche est utilisée sans connaissance de l'autre. Ensuite, chaque modèle représentant un compromis sur le front de Pareto de GANSan, est évalué avec les métriques de FM, permettant de

construire ainsi un deuxième front de Pareto pour GANSan représentant son application dans la perspective FM. Ainsi, pour chaque compromis sur le front de Pareto GANSan, nous calculons les métriques de FM (Fid_{priv}, $BER_{og_{prv}}$ et Pc_{prot}) puis nous filtrons les résultats afin de construire un nouveau front de Pareto basé sur l'ensemble de métriques de FM. De même, nous évaluons FM avec l'ensemble de métriques de GANSan et construisons un deuxième front de Pareto pour FM.

Toutes ces approches sont comparées à des références, qui correspondent aux données d'origine sans aucune modification.

Classifieurs externes. Le calcul de certaines métriques repose sur la prédiction faite par des classifieurs, tout comme dans GANSan. Nous calculons ces métriques avec un ensemble de classifieurs appelés classifieurs externes car ils sont indépendants du cadre de protection de l'attribut sensible. Pour la comparaison de l'état de l'art, l'ensemble de classifieurs externes utilisé est composé de *Gradient Boosting Classifier* (GBC) et des machines à vecteur de supports *Support Vector Machine (SVM)*. Pour chaque métrique, nous rapportons uniquement le meilleur résultat obtenu parmi tous les classifieurs externes, ce qui correspond à la pire valeur pour la métrique $classification$ (le pire cas possible) ; la protection la plus faible possible (le BER le plus élevé et le $SAcc$ le plus bas).

Tous les classifieurs externes sont obtenus par la librairie *scikit-learn*[1]. Malgré le fait que nous ayons essayé un ensemble diversifié de classifieurs, il pourrait exister un classifieur (ou un ensemble d'hyperparamètres qui conduirait à un classifieur) avec un pouvoir prédictif plus élevé pour déduire l'attribut sensible de notre transformation.

4.2.2 Résultats de la comparaison avec l'état de l'art

Dans cette section, nous présentons les résultats obtenus en utilisant les données transportées du groupe protégé et les données reconstruites du groupe privilégié. Cette utilisation correspond à celle où le modèle G_w est utilisé sans la connaissance de l'attribut sensible, ainsi, nous ne pouvons pas identifier le groupe privilégié pour décider si nous devons appliquer ou non le processus de transformation sur ces données. Nous avons présenté dans le tableau 1.3 (partie 1.5) les valeurs optimales des métriques de protection.

1. https ://scikit-learn.org/stable/

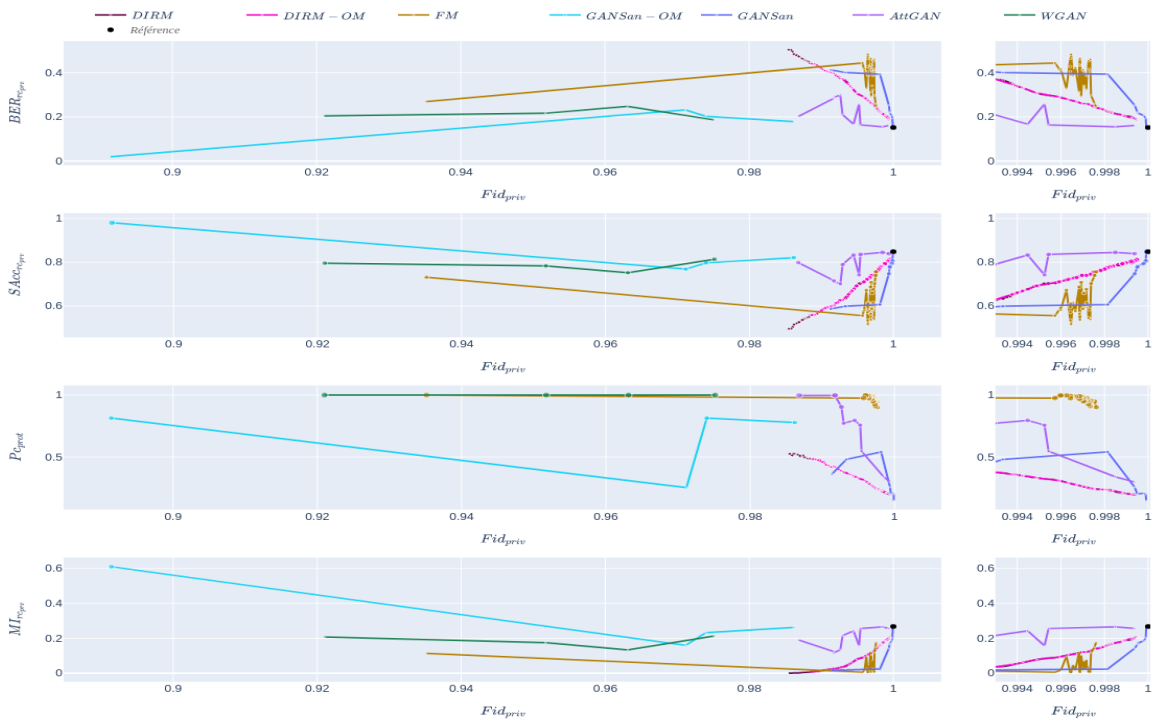


FIGURE 4.4 – Fronts de Pareto sur l'ensemble de données Lipton construit à partir des métriques de *Fair Mapping*. La colonne de gauche présente toutes les solutions sur les fronts, tandis que les colonnes de droite présentent les mêmes résultats mais dans l'intervalle $[0,985-1]$, pour une meilleure visualisation. De haut en bas : protection $BER_{rc_{prv}}$, protection $SAcc_{rc_{prv}}$, précision de S, Pc_{prot} et information mutuelle $MI_{rc_{prv}}$.

Résultats sur l'ensemble de données Lipton. Dans la figure 4.4, nous présentons les résultats obtenus sur le jeu de données *Lipton*. L'ensemble de données est équilibré, et la protection optimale est obtenue avec un BER de 0.5 et une exactitude de prédiction de 0.5. Comme nous pouvons l'observer sur la figure, le pire compromis entre la protection $BER_{rc_{prv}}$ et la fidélité Fid_{priv} est obtenu avec WGAN, suivi de $GANSan - OM$. Néanmoins, WGAN maximise la métrique de transformation Pc_{prot}, contrairement à GANSan-OM. Bien que la faible protection offerte par GANSan-OM soit inattendue, sa performance élevée sur la classification peut être expliquée par la reconstruction du groupe protégé, limitant ainsi la transformation vers le groupe privilégié. En effet, en observant les résultats de GANSan et de DIRMoù il n'y a pas de transformation vers le groupe privilégié, Pc_{prot} est proche d'une estimation aléatoire.

AttGAN réussit à transformer les données vers le groupe privilégié avec un impact léger sur la fidélité. Cependant, les protections mesurées ne sont pas parmi les meilleures possibles. DIRM et DIRM-OM obtiennent les meilleurs niveaux de protection parmi toutes les approches, mais les plus faibles transformations. DIRM-OM se comporte de manière similaire à DIRM, même si elles optimisent différents objectifs. Cela peut s'expliquer par le fait qu'elles reposent sur la même procédure sous-jacente de construction de la fonction de répartition médiane. GANSan domine toutes les approches en termes de protection et de fidélité (GANSan fournit la plus haute valeur de $BER_{rc_{prv}}$ pour les valeurs de fidélité Fid_{priv} supérieures à $0,998$), mais ne parvient pas à transformer les données vers la distribution privilégiée, laissant ainsi les données sur une distribution médiane non contrôlée.

Notre approche FM fournit le meilleur compromis sur toutes les métriques. La transformation Pc_{prot} est proche de 1, tandis que la protection est bien supérieure aux autres approches lorsque l'on considère la plage de fidélité de $[0.995 - 0.998]$. Nous pouvons également observer que l'information mutuelle est proche de zéro pour la plupart de nos résultats sur le front de Pareto.

Résultats sur German Credit. Sur German credit (figure 4.5), le groupe protégé ne représente que 19% du jeu de données. En tant que tel, le BER de prédiction de l'attribut sensible est déjà proche de 0.3 sur les données originales, ce qui rend la protection de l'attribut sensible plus difficile. En fait, DIRM et GANSan, qui présentent certaines des meilleures performances sur Lipton, n'obtiennent pas des performances (mesurées par le BER) au delà de la référence $BER_{rc_{prv}} = 0,3$. AttGAN ne permets pas d'obtenir de meilleurs résultats. WGAN obtient de meilleurs résultats que les autres approches en termes de protection, mais pour un coût important sur la fidélité Fid_{priv} (la plus haute fidélité atteignable avec l'approche est inférieure à 0.8). Comme on pouvait s'y attendre de la part de WGAN et AttGAN sur les propriétés de transformation, ces approches transforment presque parfaitement les données du groupe protégé de sorte qu'elles soient prédites comme faisant partie de la distribution privilégiée. FM domine toutes les approches sur presque toutes les métriques, et permet d'obtenir des résultats proches des meilleurs atteignables sur les autres métriques où l'approche n'est pas la meilleure. Les performances de FM peuvent être expliquées par deux facteurs : l'utilisation de l'information mutuelle et le transport des données protégées vers la distribution privilégiée. D'une part, la minimisation de l'information mutuelle encourage l'approche à mieux protéger l'attribut sensible en éliminant davantage la corrélation due à la taille du groupe. D'autre part, le transport du groupe protégé vers le groupe privilégié offre l'avantage de réduire la complexité

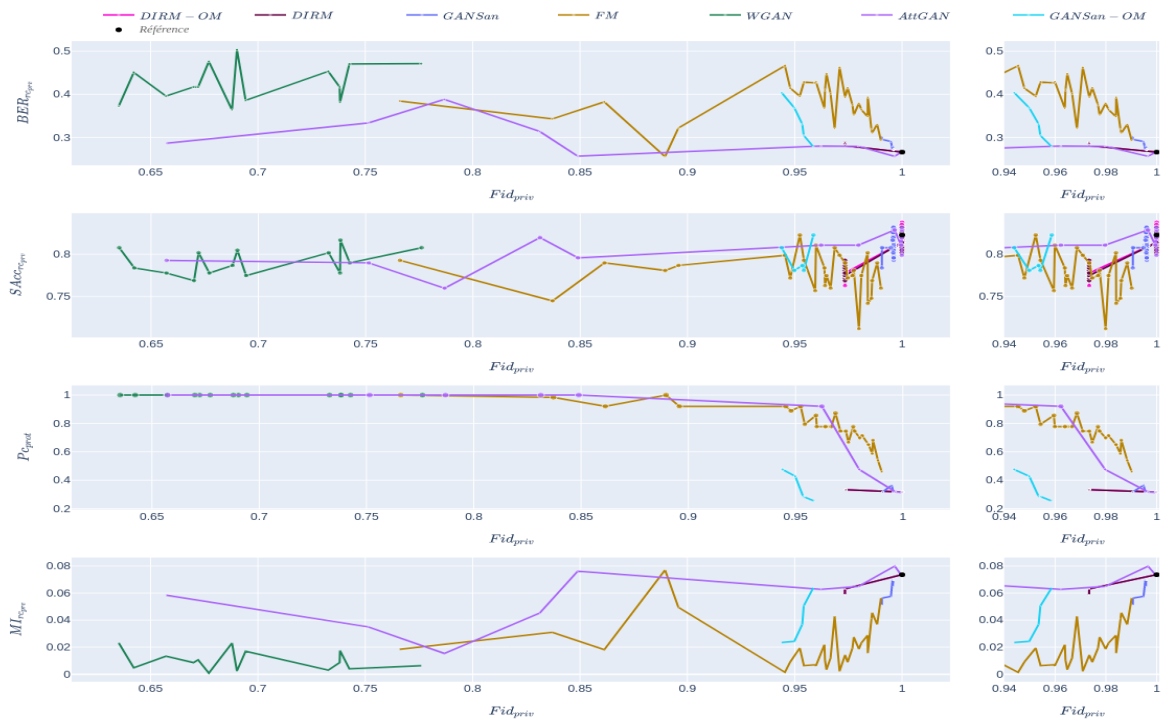


FIGURE 4.5 – Fronts de Pareto sur l'ensemble de données German Credit construit à partir des métriques de *Fair Mapping*. La colonne de gauche présente toutes les solutions sur les fronts, tandis que les colonnes de droite présentent les mêmes résultats dans l'intervalle [0.94−1], pour une meilleure visualisation. De haut en bas : protection $BER_{rc_{prv}}$, protection $SAcc_{rc_{prv}}$, précision de S, Pc_{prot}, information mutuelle $MI_{rc_{prv}}$.

de la recherche de la distribution intermédiaire idéale qui protégerait l'information sensible, en plus de ne pas modifier le groupe largement présent dans l'ensemble de données (puisque le groupe privilégié est largement présent dans l'ensemble de données, les modèles doivent seulement trouver la modification appropriée des groupes protégés). GANSan-OM améliore la protection, mais ne parvient toujours pas à transformer les données ($Pc_{prot} \leq 0.5$). DIRM-OM se comporte de manière identique à DIRM.

Résultats sur l'ensemble de données Adult Census. Les résultats sur l'ensemble de données Adult (Figure 4.6) montrent que la protection de l'attribut sensible peut être difficile à obtenir sur des distributions complexes. En fait, les approches plus simples telles que DIRM et DIRM-OM n'améliorent pas les performances sur l'ensemble de données original, à savoir $BER_{rc_{prv}} = 0.16$ et

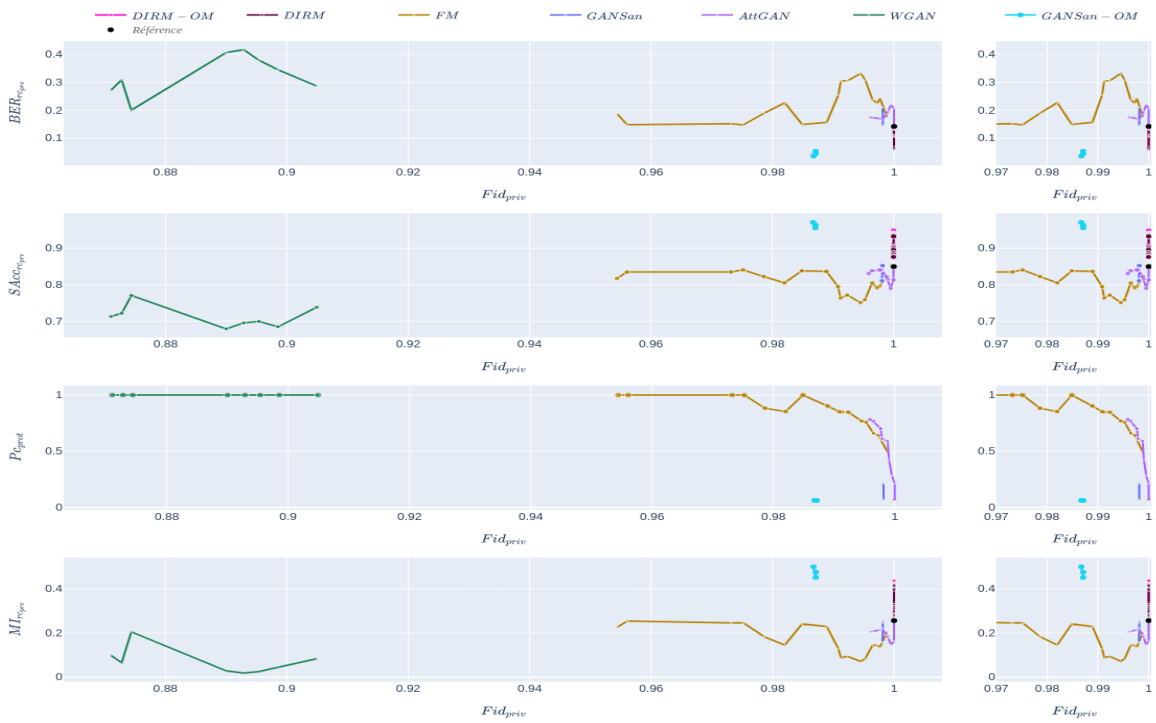


FIGURE 4.6 – Fronts de Pareto sur l'ensemble de données Adult construit à partir des métriques de *Fair Mapping*. La colonne de gauche présente toutes les solutions sur les fronts, tandis que les colonnes de droite présentent les mêmes résultats, mais sur la plage $[0,99-1]$, pour une meilleure visualisation.

$SAcc_{rc_{prv}} = 0.85$. Dans certains cas, les modifications introduites par ces approches rendent plus facile l'inférence de l'attribut sensible. GANSan-OM détériore également les résultats. Ces observations suggèrent que la reconstruction du groupe protégé (rappelons que GANSan-OM doit protéger l'attribut sensible en modifiant de manière limitée uniquement les données protégées) entrave la qualité de la protection, et empêche la modification appropriée des données afin de protéger l'attribut sensible.

Les approches GANSan et AttGAN permettent d'atteindre des fidélités parmi les plus élevées, au-delà de 0.997. Les meilleurs compromis entre la protection de l'attribut sensible et la fidélité sont atteints avec AttGAN lorsque l'on considère les valeurs de fidélité très grandes (supérieure à 0.997). Notre approche FM domine toutes les autres approches, à l'exception de WGAN, en offrant les meilleures protections pour des valeurs de fidélité au-delà de 0.98. La protection maximale mesurée

par le BER est de 0.33, correspondant au double de la valeur sur l'ensemble de données original (0.16). Pour ce compromis, la transformation Pc_{prot} se situe autour de $0,77$. WGAN présente la plus grande protection possible parmi toutes les approches. Ce niveau de protection a un coût important sur la fidélité. Néanmoins, WGAN est la seule approche permettant d'atteindre la transformation parfaite ($Pc_{prot} = 1$) pour tous les résultats de son front de Pareto.

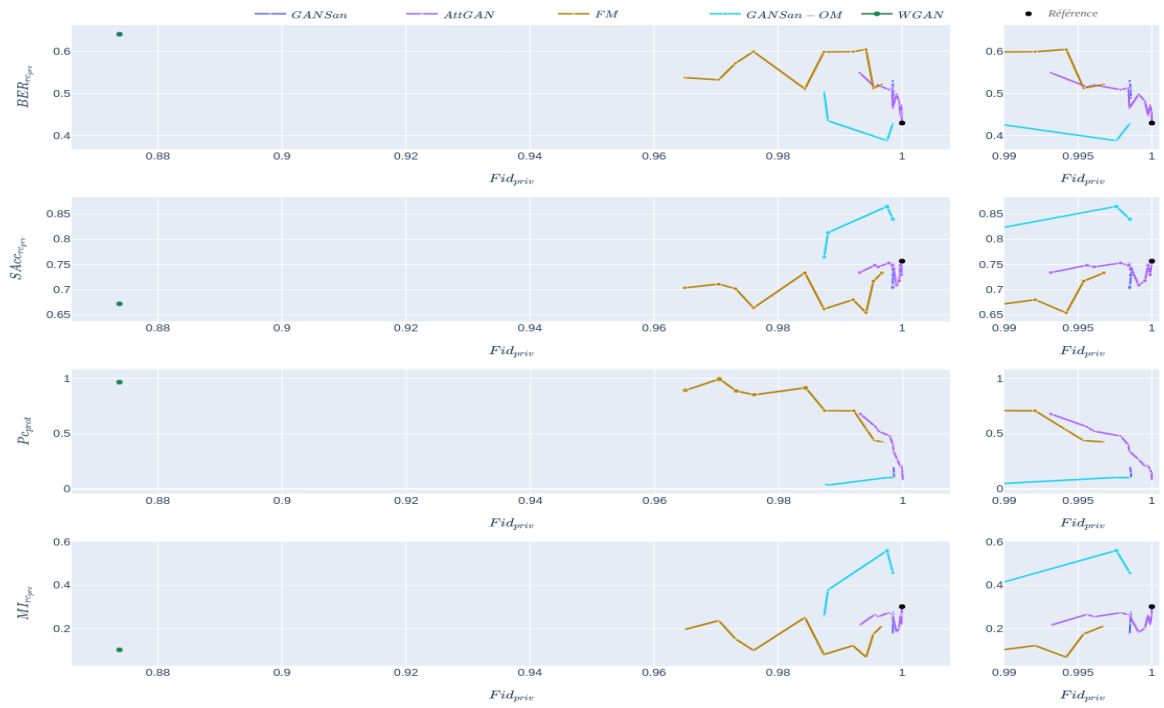


FIGURE 4.7 – Fronts de Pareto sur l'ensemble de données Adult avec 2 attributs, construit à partir des métriques de *Fair Mapping*. La colonne de droite présente toutes les solutions sur les fronts, tandis que les colonnes de droite présentent les mêmes résultats, mais dans l'intervalle $[0,99$ - $1]$, pour une meilleure visualisation. DIRM et DIRM-OM ne sont pas représentées car ces approches ne sont pas adaptées au cas multi-attribut.

Résultats sur Adult avec 2 attributs sensibles. Nous présentons les résultats de la protection de plus de 2 groupes sur l'ensemble de données Adult Census (*Adult2*) en Figure 4.7. Notez que la meilleure protection est obtenue à une valeur de BER de $\frac{3}{4}$, un $SAcc$ de 0.6.

La protection avec plus d'un seul attribut est plus difficile à réaliser pour toutes les approches. Notre approche FairMapping est capable de protéger l'attribut sensible tout en améliorant la transforma-

tion du groupe protégé. Nous pouvons observer que seules WGAN et FM sont capables d'obtenir une classification Pc_{prot} supérieure à 0.5. De même, les deux approches atteignent un niveau d'information mutuelle similaire, nettement inférieur à la référence obtenue sur l'ensemble de données original ($MI \approx 0.3$). Comme on l'a observé sur d'autres ensembles de données, GANSan-OM est incapable d'améliorer les performances des métriques au-delà de la référence.

La plus haute protection de FM mesurée avec $BER_{rc_{prv}}$ est atteinte à la valeur de $0,632$. À ce stade, le $SAcc_{rc_{prv}}$ mesuré est de $0,67$, la fidélité est de $0,9946$ et la transformation a la valeur de $Pc_{prot} = 0,662$. En plus de ce point, il existe un autre dont la plus haute protection mesurée avec le $SAcc_{rc_{prv}}$ a un meilleur résultat de transformation, mais présente une légère diminution de la fidélité : $BER_{rc_{prv}} = 0,63$, $SAcc_{rc_{prv}} = 0,65$, $Pc_{prot} = 0,72$, $Fid_{prv} = 0,9930$.

Nous pouvons également expliquer nos résultats par le fait que nous avons utilisé la structure de modèle qui maximise la fidélité. En tant que tel, il pourrait être possible d'améliorer encore les résultats en relaxant les contraintes sur la fidélité et en utilisant une autre structure de modèle qui favoriserait moins cette métrique. Cependant, cette relaxation augmenterait l'espace des hyperparamètres et pourrait entraîner une augmentation de la durée d'entraînement et de recherche d'hyperparamètres.

Divergences. Comme FairMapping transporte les données protégées sur la distribution privilégiée, nous mesurons la proximité entre la distribution protégée transformée et celle du groupe privilégié. Nous nous basons sur la divergence *Sinkhorn* (Cuturi, 2013) pour calculer la divergence entre le groupe privilégié original et les données protégées transformées ($R_{priv} \sim G_w(R_{prot})$), entre le groupe privilégié reconstruit (obtenu par G_w) et le groupe protégé transformé ($G_w(R_{priv}) \sim G_w(R_{prot})$) et enfin entre les données protégées d'origine et leur version transformée ($R_{prot} \sim G_w(R_{prot})$). L'objectif ici est d'obtenir $R_{priv} \sim G_w(R_{prot})$ et $G_w(R_{priv}) \sim G_w(R_{prot})$ inférieur à $R_{prot} \sim G_w(R_{prot})$. Idéalement, nous souhaitons obtenir la valeur de 0 pour $R_{priv} \sim G_w(R_{prot})$ et $G_w(R_{priv}) \sim G_w(R_{prot})$. Les résultats sont présentés en figure 4.8.

Nous pouvons observer que, plus la protection est élevée, plus la valeur des divergences de $G_w(R_{priv}) \sim G_w(R_{prot})$ est petite, et haute est celle de $R_{prot} \sim G_w(R_{prot})$. Notre approche est capable d'obtenir les résultats de divergence le plus bas lorsque les protections sont plus élevées. Nous pouvons également observer que la divergence *Sinkhorn* entre les données protégées originales et leur version transformée ($R_{prot} \sim G_w(R_{prot})$) diminue lorsque la fidélité augmente. Néanmoins, ces valeurs de

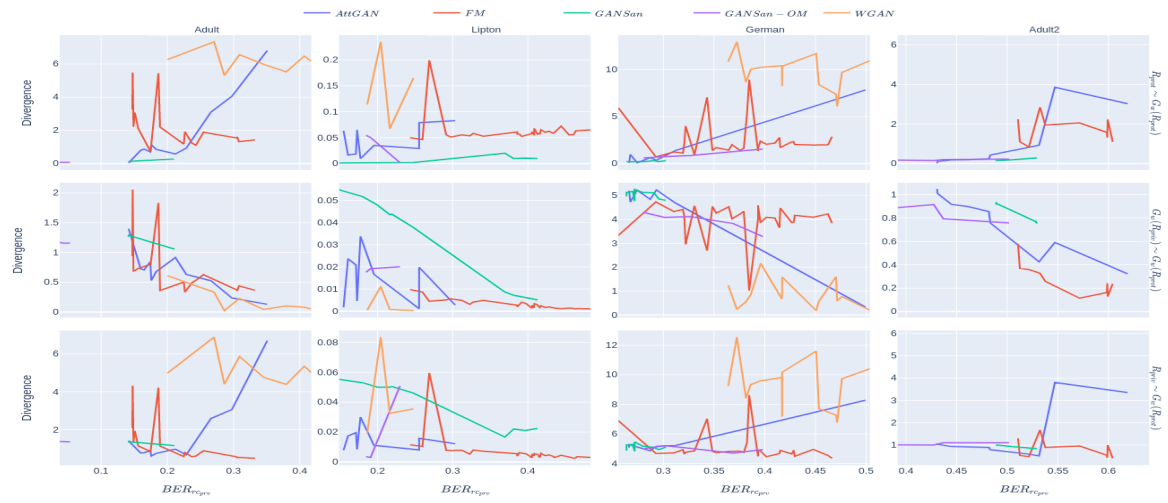


FIGURE 4.8 – Divergences obtenies sur German, Adult, Adult2 et Lipton. Chaque colonne représente un ensemble de données tandis que chaque ligne représente la divergence entre les données protégées transformées et respectivement les données protégées originales ($R_{prot} \sim G_w(R_{prot})$), le groupe privilégié reconstruit ($G_w(R_{priv}) \sim G_w(R_{prot})$), et le groupe privilégié original ($R_{priv} \sim G_w(R_{prot})$). Les divergences sont représentées par rapport à la protection $BER_{rc_{prv}}$.

divergence restent supérieures aux divergences calculées avec le groupe privilégié ($R_{priv} \sim G_w(R_{prot})$ et $G_w(R_{priv}) \sim G_w(R_{prot})$). Cela suggère qu'à mesure que la fidélité augmente, notre modèle de FM est capable d'apprendre la structure des données et de produire un ensemble de données avec moins de valeurs aléatoires.

Nous pouvons également observer que la distribution intermédiaire obtenue avec d'autres approches de prétraitement (GANSan etAttGAN) est plus proche de la distribution protégée, mais très éloignée de la distribution privilégiée. Ces résultats sont surprenants, notamment sur l'ensemble de données de German credit, où le groupe privilégié est largement présent. Avec FairMapping, la distribution transformée est plus proche de la distribution du groupe privilégié. Par conséquent, un classifieur entraîné sur les données originales pourrait considérer plus facilement le profil transformé obtenu avec FairMapping comme appartenant à la même frontière de décision que le groupe privilégié.

Instances de FairMapping. Comme mentionné dans la section 4.1, notre approche peut être instanciée avec le BER ou avec l'exactitude de prédiction, avec ou sans la régularisation de l'information mutuelle. Les résultats de ces différentes instantiatons sont présentées dans la figure 4.9.

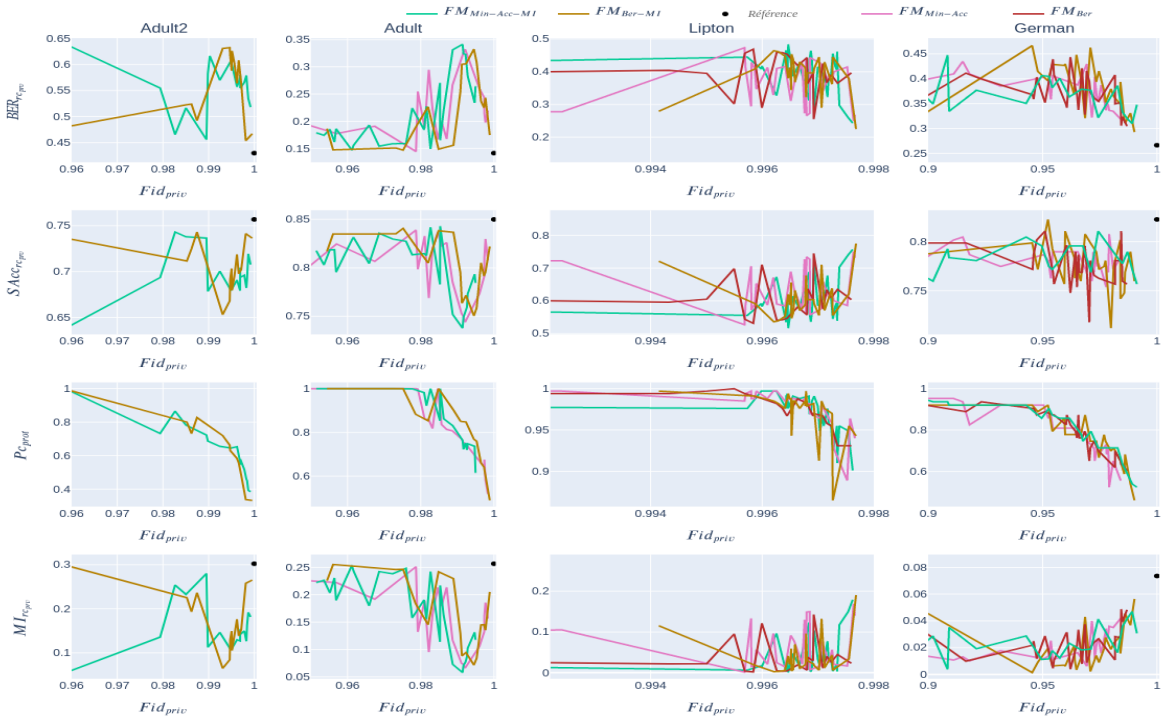


FIGURE 4.9 – Différentes instances de *Fair Mapping*. Nous pouvons observer que les instances FM se comportent de manière à peu près similaire pour toutes les fonctions de perte choisies pour protéger l'attribut sensible. Chaque colonne représente les performances sur un seul jeu de données. Sur Adult2, nous incluons toujours la régularisation de l'information mutuelle.

Nous pouvons observer que les différentes instances de FM se comportent de manière presque similaire sur tous les jeux de données de notre expérimentation, l'information mutuelle améliore légèrement les résultats là où elle est introduite. Sur le jeu de données German credit, nous pouvons observer que la régularisation par information mutuelle aide à améliorer les BER (tous les FM qui utilisent la MI sont légèrement meilleurs que ceux qui ne l'utilisent pas). Sur Lipton, l'ensemble de données équilibré, l'information mutuelle n'a pas d'impact significatif, contrairement à ce qui a été observé pour German Credit. Cela suggère que dans une situation où le jeu de données est fortement déséquilibré, l'introduction de la régularisation de l'information mutuelle améliorerait la qualité des résultats.

4.2.3 Temps d'exécution

Nous présentons dans le tableau 4.3 le temps d'exécution moyen pour chaque approche pour compléter une époque (une boucle sur l'ensemble des données d'entraînement) de calcul sur l'ensemble de données Lipton. L'approche DIRM prend moins de 10 secondes pour terminer, elle n'est donc pas représentée dans le tableau. Notre approche FM est la plus longue, tandis que GANSan est la

Tableau 4.3 – Temps d'exécution moyen calculé sur le jeu de données Lipton, calculé avec un seul GPU

Approche	FM	FM Avec MI	GANSan	GANSan-OM	AttGAN	WGAN
Temps par époque	$0,224337$	$0,236976$	$0,126364$	$0,140211$	$0,141715$	$0,183009$
Temps pour 1000 époques *(mm :ss)*	$0:03:44$	$0:03:56$	$02:06$	$02:20$	$02:21$	$03:03$

plus rapide à l'exception de DIRM. Cela est dû au fait que notre approche a le plus grand nombre de modèles impliqués dans le calcul. Le temps d'exécution est amélioré par le fait que le classifieur C peut être entraîné indépendamment des modèles principaux de notre approche (G_w, D et D_{std}).

Lors de nos expériences avec l'ensemble de données *Adult*, nous avons constaté que sur un GPU disposant de 4 Go de mémoire, nous pouvons faire fonctionner simultanément jusqu'à 3 instances différentes de FM avec un peu de mémoire résiduelle, 4 instances de GANSan, GANSan-OM et WGAN, ainsi que 2 instances de AttGAN. En conséquence, l'exploration des hyperparamètres peut être un peu plus rapide pour FM.

4.3 FM : conclusion

En résumé, notre approche FairMapping consiste en la protection de l'attribut sensible par la transformation d'un profil de sorte que ce dernier appartient à une distribution cible déterminée. Cette approche préserve notamment l'aspect réaliste des données tout en permettant aux données transformées de bénéficier des mêmes avantages ou privilèges (et aussi inconvénients) présents dans le groupe cible. Le transport vers une distribution connue permet aussi d'éviter les complexités dues au réentraînement de modèles déjà existants (comme c'est le cas pour la plupart des approches de prétraitement des données), puisque ces modèles ont déjà été entraînés sur la distribution cible. Notre approche peut être utilisée pour étendre certaines méthodes de l'état de l'art et nos résultats expérimentaux démontrent que l'approche et certaines extensions permettent de prévenir l'inférence

de l'attribut sensible, avec un coût très avantageux sur les données du groupe cible choisi. Toutefois, le processus d'apprentissage de FairMapping peut être relativement complexe, particulièrement avec les données catégoriques et d'autres qui seraient corrélées à la fois avec l'attribut sensible et l'attribut de décision.

Bien que nous n'ayons pas pu évaluer les capacités de notre approche pour la situation de multiples attributs sensibles, notre cadriciel et notre approche intègrent automatiquement cette possibilité. Une telle ouverture constituerait une prochaine direction de recherche. De même, FairMapping offre aussi la possibilité de travailler sur des groupes inconnus durant la phase d'entraînement, mais dont l'existence pourrait être obtenue par la combinaison de valeurs des attributs sensibles. Par exemple, le groupe inconnu *Femme-Noire* pourrait être construit à partir des groupes connus *Homme-Noir* et *Femme-Blanche*. Nos travaux futurs suivront ces différentes orientations.

CONCLUSION

L'équité en apprentissage machine est un domaine très vaste et qui commence à peine à être exploré. Plusieurs des problématiques dans le domaine proviennent d'enjeux sociaux et sociétaux comme nous l'avons vu dans notre introduction. Dans le premier chapitre, nous avons présenté un tour d'horizon de l'origine des biais dans les systèmes automatiques, discuté des approches et techniques utilisées pour l'amélioration de l'équité et discuté des relations avec d'autres domaines tels que la protection de la vie privée et l'optimisation robuste. Nous avons aussi vu que plusieurs techniques se concentrent sur l'amélioration de l'équité de groupe et que beaucoup d'aspects tels que l'impact des métriques et des méthodes sont très peu explorés. Nous avons aussi proposé une classification afin de faciliter l'identification des approches, pour permettre aux nouvelles personnes dans le domaine de s'y retrouver facilement et de choisir les approches les mieux adaptées à leur problème.

Dans le second chapitre, nous avons introduit notre première approche d'amélioration de l'équité, GANSan. Cette approche est inspirée des réseaux adversariaux génératifs et modifie les données dans l'objectif d'améliorer l'équité en rendant toute information concernant l'attribut sensible non disponible à partir du jeu de donnée ainsi prétraité. Par conséquent, toute utilisation faite du jeu de données est libre de discrimination puisque l'information sensible à partir de laquelle on peut discriminer est indisponible et irrécupérable. Nous avons montré comment l'utilisation de l'approche permet d'améliorer les prédictions dans le groupe et caractérisé le comportement de notre solution au travers de différentes compositions du jeu de données. Dans cette approche, nous avons aussi introduit différents scénarii d'évaluation qui ont inspiré le critère *scénario* dans notre premier chapitre et montré comment notre approche de prétraitement renforce la protection de la vie privée en rendant le contrôle de l'information à l'utilisateur, notamment au travers du scénario d'assainissement local.

Comme présenté dans le premier chapitre, le domaine de l'équité est en étroite relation avec d'autres. Ceci est notamment le cas de notre troisième chapitre, dans lequel nous avons introduit la méthode de protection des données de capteurs DYSAN. Cette approche correspond à une amélioration et une extension de notre approche GANSan dans le contexte de protection des données de santé. En particulier, nous avons étendu l'approche pour montrer comment la protection de plusieurs personnes

est affectée par l'utilisation de modèles uniques qui ont des performances moyennes. La généralisation peut être très difficile, étant donné que la minimisation de l'erreur empirique se fait sur la moyenne de l'erreur calculée sur tous les profils de l'ensemble de données. Notre approche est développée avec le maintien dynamique de sélection des modèles pour maximiser l'utilité et la protection des informations privée.

Enfin, dans notre dernier chapitre, l'approche *FairMapping* est présentée. Celle-ci a pour objectif de pallier les lacunes des approches de prétraitement des données, notamment le choix de la distribution cible et le maintien de l'aspect réaliste de la transformation des données. L'approche permet aussi de gérer plusieurs attributs dans le prétraitement, permet une réutilisation des données sans nécessiter un réentraînement des modèles existants, confère à tous les groupes les mêmes avantages qu'un groupe choisi, ne nécessite pas de connaissance de l'attribut sensible au niveau du déploiement du modèle, et peut aussi être utilisée dans le contexte de la détection de discrimination.

Nous avons développé dans nos travaux de thèse des approches basées les RAG pour améliorer la prise de décision équitable par des modèles automatiques. Bien que nous ayons pu montrer l'utilité de nos méthodes, plusieurs approfondissements restent nécessaires. Dans un premier temps, avec notre approche *GANSan*, nous avons pu montrer que la décision a un impact non négligeable sur la protection de l'attribut sensible, puisque celle-ci peut être biaisée. Il devient ainsi difficile de déterminer en avance la décision correcte à utiliser durant l'entraînement des modèles de prédictions : l'utilisation de la décision originale est justifiée par le fait qu'il s'agisse d'une décision réelle bien qu'elle soit biaisée alors que l'utilisation de la décision assainie est justifiée par le fait que celle-ci est dépourvue de toute discrimination. Cependant, non seulement elle peut ne pas être réaliste, mais plusieurs décisions assainies différentes peuvent être obtenues selon la méthode de prétraitement des données utilisée, rendant le choix de la décision encore plus complexe. Le choix de la décision à utiliser reste ainsi difficile, car il n'existe pas de critère sûr permettant d'orienter ce choix.

De plus, la majorité des méthodes de prétraitement des données développées pour l'amélioration de l'équité ne tiennent pas compte de la *situation partiellement informée* dans laquelle les attributs sensibles sont partiellement connus, et ne sont pas non plus développées avec la capacité de gérer plusieurs attributs sensibles. Ces limites ne permettent pas en conséquence d'évaluer le *glissement de la discrimination*, où l'amélioration de l'équité mesurée par rapport à un premier attribut conduit à une augmentation de la discrimination sur un autre. L'analyse de ce glissement aussi importante dans la mesure où la plupart des ensembles de données contiennent plus d'un attribut sensible, et il

serait nécessaire de décider en amont si l'amélioration de l'équité doit être faite en considérant tous les attributs sensibles ou en considérant uniquement un sous ensembles de ces attributs.

Ensuite, il serait important d'évaluer les performances des méthodes de prétraitement des données dans le cadre de leur utilisation avec des classifieurs entraînés avec des mécanismes de robustesse. En effet, les classifieurs peuvent-être sujet à des attaques d'adversaires, dont l'objectif est de générer des exemples contradictoires, c'est-à-dire des intrants du modèle délibérément conçus pour que le modèle fasse des erreurs dans ses prédictions, même s'ils ressemblent à des intrants valides pour un humain. En conséquence, des techniques ont été conçues pour rendre le modèle de prédiction invariant aux changements générés dans les données par ce type d'attaque. Les techniques de prétraitement qui introduisent des perturbations dans les données peuvent ainsi avoir des performances très impactées par les mécanismes de protection contre les attaques.

Une autre direction de recherche nécessaire à l'amélioration de l'équité de la prise de décisions automatique est la gestion des valeurs manquantes. Cette direction est notamment plus importante dans la mesure où la sensibilité des personnes aux différents attributs peut être différente. Ainsi, avec l'amélioration des techniques de protection des données personnelles, on pourrait se retrouver avec des ensembles de données contenant des valeurs manquantes dont l'inférence est rendue difficile par ces mécanismes. Très peu de recherche étudient cet aspect. Celles qui évaluent l'impact des données manquantes sur l'équité des décisions considèrent que celles-ci représentent une faible proportion de l'ensemble de données, et ne considère pas que les attributs sensibles peuvent-être définie par chaque personne selon ses sensibilités.

Pour terminer, l'équité dans l'apprentissage automatique est un domaine en évolution, et doit pouvoir intégrer des connaissances plus larges de la société et des problèmes sociaux. Spécifiquement, ce domaine doit pouvoir être adapté à différentes communautés et cultures, et doit permettre l'ouverture aux autres, à la différence et surtout la prise de conscience que peuvent avoir les discriminations et les décisions injustes sur les vies des personnes humaines. En effet, une fois la dignité d'un être humain affectée, celle-ci est difficilement reconstruite (Martin, 2019).

RÉFÉRENCES

Aalmoes, Jan, Vasisht Duddu et Antoine Boutet (fév. 2022). « Dikaios : Privacy Auditing of Algorithmic Fairness via Attribute Inference Attacks ». In : *arXiv preprint arXiv :2202.02242* abs/2202.02242.

Abadi, Martin, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar et Li Zhang (oct. 2016). « Deep learning with differential privacy ». In : *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. Sous la dir. d'Edgar R. Weippl, Stefan Katzenbeisser 0001, Christopher Kruegel, Andrew C. Myers et Shai Halevi. ACM, p. 308-318. doi : 10.1145/2976749.2978318.

Abadleh, A., E. Al-Hawari, E. Alkafaween et H. Al-Sawalqah (mai 2017). « Step detection algorithm for accurate distance estimation using dynamic step length ». In : *MDM*. T. abs/1801.02336. IEEE, p. 324-327. doi : 10.1109/mdm.2017.52.

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudík, John Langford et Hanna Wallach (mars 2018). « A reductions approach to fair classification ». In : *International Conference on Machine Learning*. Sous la dir. de Jennifer G. Dy et Andreas Krause 0001. T. 80. PMLR. JMLR.org, p. 60-69.

Agarwal, Pankaj K (nov. 2018). « Public administration challenges in the world of AI and bots ». In : *Public Administration Review* 78.6, p. 917-921. issn : 0033-3352.

Agarwal, Sushant (2020). « Trade-offs between fairness, interpretability, and privacy in machine learning ». Mém. de mast. University of Waterloo.

Aïvodji, Ulrich, François Bidet, Sébastien Gambs, Rosin Claude Ngueveu et Alain Tapp (mars 2021). « Local Data Debiasing for Fairness Based on Generative Adversarial Training ». In : *Algorithms* 14.3, p. 87. issn : 1999-4893.

Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta et Masanori Koyama (2019). « Optuna : A next-generation hyperparameter optimization framework ». In : *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, p. 2623-2631.

Alemdar, Hande, Can Tunca et Cem Ersoy (fév. 2015). « Daily life behaviour monitoring for health assessment using machine learning : bridging the gap between domains ». In : *Personal and Ubiquitous Computing* 19.2, p. 303-315. issn : 1617-4909. doi : 10.1007/s00779-014-0823-y.

Anastasopoulos, L Jason et Andrew B Whitford (juin 2019). « Machine learning for public administration research, with application to organizational reputation ». In : *Journal of Public Administration Research and Theory* 29.3, p. 491-510. issn : 1053-1858. doi : 10.1093/jopart/muy060.

Angwin, Julia, Jeff Larson, Surya Mattu et Lauren Kirchner (2016). *Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks.* (Visité le 14/02/2022).

Arjovsky, Martin, Soumith Chintala et Léon Bottou (2017). « Wasserstein generative adversarial networks ». In : *International conference on machine learning*. Sous la dir. de Doina Precup et Yee Whye Teh. T. 70. PMLR. JMLR.org, p. 214-223.

Arsenault, Marie-Eve (2022). *Les agents correctionnels ont plus souvent recours à la force face aux Autochtones.* (Visité le 11/02/2022).

Ayodele, Taiwo Oladipupo (2010). « Types of machine learning algorithms ». In : *New advances in machine learning* 3, p. 19-48.

Bagdasaryan, Eugene, Omid Poursaeed et Vitaly Shmatikov (mai 2019). « Differential privacy has disparate impact on model accuracy ». In : *Advances in neural information processing systems* 32. Sous la dir. d'Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Edward A. Fox et Roman Garnett, p. 15453-15462.

Balashankar, Ananth, Alyssa Lees, Chris Welty et Lakshminarayanan Subramanian (2019). « Pareto-Efficient Fairness for Skewed Subgroup Data ». In : *International Conference on Machine Learning AI for Social Good Workshop. Long Beach, United States*. T. 8.

Barker, Ken, Mina Askari, Mishtu Banerjee, Kambiz Ghazinour, Brenan Mackas, Maryam Majedi, Sampson Pun et Adepele Williams (2009). « A data privacy taxonomy ». In : *British National Conference on Databases*. Sous la dir. d'Alan P. Sexton. T. 5588. Springer. Springer Berlin Heidelberg, p. 42-54. isbn : 9783642028427. doi : 10.1007/978-3-642-02843-4_7.

Barocas, Solon, Moritz Hardt et Arvind Narayanan (2019). *Fairness and Machine Learning*. http ://www.fairmlbook.org. fairmlbook.org.

Barocas, Solon et Andrew D Selbst (2016). « Big data's disparate impact ». In : *Cal. L. Rev.* 104, p. 671. issn : 1556-5068. doi : 10.2139/ssrn.2477899.

Bechavod, Yahav, Christopher Jung et Steven Z Wu (2020). « Metric-free individual fairness in online learning ». In : *Advances in neural information processing systems* 33. Sous la dir. d'Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan et Hsuan-Tien Lin, p. 11214-11225.

Bellamy, Rachel KE, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic et al. (oct. 2018). « AI Fairness 360 : An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias ». In : *arXiv preprint arXiv :1810.01943* abs/1810.01943.

BenAbdelkader, C., R. Cutler et L. Davis (juin 2002). « Stride and cadence as a biometric in automatic person identification and verification ». In : *IEEE FGR*. IEEE, p. 372-377. doi : 10.1109/afgr.2002.1004182.

Bendick, Marc (sept. 2007). « Situation testing for employment discrimination in the United States of America ». In : *Horizons stratégiques* 3.5, p. 17-39. issn : 1958-3370. doi : 10.3917/hori.005.0017.

Benussi, Elias, Andrea Patane, Matthew Wicker, Luca Laurenti et Marta Kwiatkowska (juill. 2022). « Individual Fairness Guarantees for Neural Networks ». In : *arXiv preprint arXiv :2205.05763*, p. 651-658. doi : 10.24963/ijcai.2022/92.

Berndt, Donald J. et James Clifford (1994). « Using Dynamic Time Warping to Find Patterns in Time Series ». In : *AAAIWS*, p. 359-370.

Black, Emily, Samuel Yeom et Matt Fredrikson (juin 2020). « FlipTest : fairness testing via optimal transport ». In : *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Sous la dir. de Mireille Hildebrandt, Carlos Castillo 0001, Elisa Celis, Salvatore Ruggieri, Linnet Taylor et Gabriela Zanfir-Fortuna. ACM, p. 111-121. doi : 10.1145/3351095.3372845.

Blyth, Colin R (juin 1972). « On Simpson's paradox and the sure-thing principle ». In : *Journal of the American Statistical Association* 67.338, p. 364-366. issn : 0162-1459. doi : 10.1080/01621459.1972.10482387.

Boutet, Antoine, Carole Frindel, Sébastien Gambs, Théo Jourdan et Rosin Claude Ngueveu (mai 2021). « DYSAN : Dynamically sanitizing motion sensor data against sensitive inferences through adversarial networks ». In : *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*. Sous la dir. de Jiannong Cao 0001, Man Ho Au, Zhiqiang Lin et Moti Yung. ACM, p. 672-686. doi : 10.1145/3433210.3453095.

Brandom, Russell (2018). *Amazon's facial recognition matched 28 members of Congress to criminal mugshots*.

Buolamwini, Joy et Timnit Gebru (2018). « Gender shades : Intersectional accuracy disparities in commercial gender classification ». In : *Conference on fairness, accountability and transparency*. Sous la dir. de Sorelle A. Friedler et Christo Wilson. T. 81. PMLR. PMLR, p. 77-91.

Carbonell, Jaime G, Ryszard S Michalski et Tom M Mitchell (1983). « An overview of machine learning ». In : *Machine learning*, p. 3-23. doi : 10.1007/978-3-662-12405-5_1.

Carey, Alycia N et Xintao Wu (avr. 2022). « The Causal Fairness Field Guide : Perspectives From Social and Formal Sciences ». In : *Frontiers in Big Data* 5, p. 892837. issn : 2624-909X.

Caton, Simon et Christian Haas (oct. 2020). « Fairness in machine learning : A survey ». In : *arXiv preprint arXiv :2010.04053* abs/2010.04053.

CENTER, PEW RESEARCH (2017). *Nearly half of Americans use digital voice assistants, mostly on their smartphones*. (Visité le 08/02/2022).

Chang, Hongyan et Reza Shokri (sept. 2021). « On the privacy risks of algorithmic fairness ». In : *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. IEEE, p. 292-303. doi : 10.1109/eurosp51992.2021.00028.

Chen, Stephen (2021). *Chinese scientists develop AI 'prosecutor' that can press its own charges*. (Visité le 08/02/2022).

Cortes, Corinna et Vladimir Vapnik (sept. 1995). « Support-vector networks ». In : *Machine learning* 20.3, p. 273-297. issn : 0885-6125. doi : 10.1007/bf00994018.

Cummings, Rachel, Varun Gupta, Dhamma Kimpara et Jamie Morgenstern (juin 2019). « On the compatibility of privacy and fairness ». In : *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. Sous la dir. de George Angelos Papadopoulos, George Samaras, Stephan Weibelzahl, Dietmar Jannach et Olga C. Santos. ACM, p. 309-315. doi : 10.1145/3314183.3323847.

Cuturi, Marco (2013). « Sinkhorn distances : Lightspeed computation of optimal transport ». In : *Advances in neural information processing systems* 26. Sous la dir. de Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani et Kilian Q. Weinberger, p. 2292-2300.

Danks, David et Alex John London (août 2017). « Algorithmic Bias in Autonomous Systems. » In : *IJCAI*. Sous la dir. de Carles Sierra. T. 17. International Joint Conferences on Artificial Intelligence Organization, p. 4691-4697. doi : 10.24963/ijcai.2017/654.

DASTIN, Jeffrey (mars 2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. DOI : 10.1201/9781003278290-44. (Visité le 11/02/2022).

DIXON, Pam (2013). « Congressional testimony : what information do data brokers have on consumers ? » In : *World Privacy Forum*. T. 15.

DRINHAUSEN, Katja et Vincent BRUSSEE (2021). *China's Social Credit System in 2021 : From fragmentation towards integration*.

DWORK, Cynthia, Moritz HARDT, Toniann PITASSI, Omer REINGOLD et Richard ZEMEL (avr. 2012). « Fairness through awareness ». In : *Proceedings of the 3rd innovations in theoretical computer science conference*. Sous la dir. de Shafi GOLDWASSER. T. abs/1104.3913. ACM. ACM Press, p. 214-226. DOI : 10.1145/2090236.2090255.

DWORK, Cynthia, Frank MCSHERRY, Kobbi NISSIM et Adam SMITH (mai 2006). « Calibrating noise to sensitivity in private data analysis ». In : *Theory of cryptography conference*. Sous la dir. de Shai HALEVI et Tal RABIN. T. 7. Springer. Journal of Privacy et Confidentiality, p. 265-284. DOI : 10.29012/jpc.v7i3.405.

DWORK, Cynthia, Aaron ROTH et al. (2014). « The algorithmic foundations of differential privacy ». In : *Foundations and Trends® in Theoretical Computer Science* 9.3–4, p. 211-407. ISSN : 1551-305X. DOI : 10.1561/9781601988195.

EDWARDS, Harrison et Amos STORKEY (nov. 2015). *Censoring Representations with an Adversary*.

FABRIS, Alessandro, Stefano MESSINA, Gianmaria SILVELLO et Gian Antonio SUSTO (fév. 2022). « Algorithmic Fairness Datasets : the Story so Far ». In : *arXiv preprint arXiv :2202.01711* abs/2202.01711.

FACEBOOK (2022). *Facebook's Five Pillars of Responsible AI*. URL : https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/ (visité le 02/09/2022).

FAZELPOUR, Sina et David DANKS (août 2021). « Algorithmic bias : Senses, sources, solutions ». In : *Philosophy Compass* 16.8, e12760. ISSN : 1747-9991.

FEAGIN, Joe R et Douglas Lee ECKBERG (août 1980). « Discrimination : Motivation, action, effects, and context ». In : *Annual Review of Sociology* 6, p. 1-20. ISSN : 0360-0572.

FELDMAN, Michael, Sorelle A FRIEDLER, John MOELLER, Carlos SCHEIDEGGER et Suresh VENKATASUBRAMANIAN (août 2015). « Certifying and removing disparate impact ». In : *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. Sous la dir. de Longbing CAO, Chengqi ZHANG, Thorsten JOACHIMS, Geoffrey I. WEBB, Dragos D. MARGINEANTU et Graham WILLIAMS. ACM, p. 259-268. DOI : 10.1145/2783258.2783311.

Ferrer, Xavier, Tom van Nuenen, Jose M Such, Mark Coté et Natalia Criado (juin 2021). « Bias and Discrimination in AI : a cross-disciplinary perspective ». In : *IEEE Technology and Society Magazine* 40.2, p. 72-80. issn : 0278-0097. doi : 10.1109/mts.2021.3056293.

Feuerriegel, Stefan, Mateusz Dolata et Gerhard Schwabe (août 2020). « Fair AI ». In : *Business & information systems engineering* 62.4, p. 379-384. issn : 2363-7005. doi : 10.1007/s12599-020-00650-3.

Fischer, Marc, Mislav Balunovic, Dana Drachsler-Cohen, Timon Gehr, Ce Zhang et Martin Vechev (2019). « Dl2 : Training and querying neural networks with logic ». In : *International Conference on Machine Learning*. Sous la dir. de Kamalika Chaudhuri et Ruslan Salakhutdinov. T. 97. PMLR. PMLR, p. 1931-1941.

Foulds, James R, Rashidul Islam, Kamrun Naher Keya et Shimei Pan (2020). « An intersectional definition of fairness ». In : *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, p. 1918-1921.

Friedler, Sorelle A, Carlos Scheidegger et Suresh Venkatasubramanian (sept. 2016). « On the (im) possibility of fairness ». In : *arXiv preprint arXiv :1609.07236* abs/1609.07236.

Friedler, Sorelle A, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton et Derek Roth (jan. 2019). « A comparative study of fairness-enhancing interventions in machine learning ». In : *Proceedings of the conference on fairness, accountability, and transparency*. ACM, p. 329-338. doi : 10.1145/3287560.3287589.

Friedman, Jerome H (fév. 2002). « Stochastic gradient boosting ». In : *Computational Statistics & Data Analysis* 38.4, p. 367-378. issn : 0167-9473. doi : 10.1016/s0167-9473(01)00065-2.

Fung, Brian (2017). *Uber settles with FTC over 'God View' and some other privacy issues.*

Galhotra, Sainyam, Yuriy Brun et Alexandra Meliou (août 2017). « Fairness testing : testing software for discrimination ». In : *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*. Sous la dir. d'Eric Bodden, Wilhelm Schäfer, Arie van Deursen et Andrea Zisman. ACM, p. 498-510. doi : 10.1145/3106237.3106277.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii et Kate Crawford (déc. 2021). « Datasheets for datasets ». In : *Communications of the ACM* 64.12, p. 86-92. issn : 0001-0782. doi : 10.1145/3458723.

Gent, Jacob W. (2017). *The Emerging Privacy Invasion from the Insurance Industry.*

Géron, Aurélien (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow : Concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc."

Gillen, Stephen, Christopher Jung, Michael Kearns et Aaron Roth (fév. 2018). « Online learning with an unknown fairness metric ». In : *Advances in neural information processing systems* 31. Sous la dir. de Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi et Roman Garnett, p. 2605-2614.

Glorot, Xavier, Antoine Bordes et Yoshua Bengio (nov. 2011). « Deep Sparse Rectifier Neural Networks ». In : *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Sous la dir. de Geoffrey Gordon, David Dunson et Miroslav Dudík. T. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA : PMLR, p. 315-323. url : https://proceedings.mlr.press/v15/glorot11a.html.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville et Yoshua Bengio (2014). « Generative adversarial nets ». In : *Advances in neural information processing systems*. Sous la dir. de Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence et Kilian Q. Weinberger, p. 2672-2680.

Google (2022). *Our Principles*. Google AI. url : https://ai.google/principles/ (visité le 02/09/2022).

Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin et Aaron C Courville (mars 2017). « Improved training of wasserstein gans ». In : *Advances in neural information processing systems*. Sous la dir. d'Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan et Roman Garnett, p. 5767-5777.

Gupta, Maya, Andrew Cotter, Mahdi Milani Fard et Serena Wang (juin 2018). « Proxy fairness ». In : *arXiv preprint arXiv :1806.11212* abs/1806.11212.

Hardt, Moritz, Eric Price, Nati Srebro et al. (oct. 2016). « Equality of opportunity in supervised learning ». In : *Advances in neural information processing systems*. Sous la dir. de Daniel D. Lee, Masashi Sugiyama, Ulrike V. Luxburg, Isabelle Guyon et Roman Garnett, p. 3315-3323.

Hashimoto, Tatsunori, Megha Srivastava, Hongseok Namkoong et Percy Liang (juin 2018). « Fairness without demographics in repeated loss minimization ». In : *International Conference on Machine Learning*. Sous la dir. de Jennifer G. Dy et Andreas Krause 0001. T. 80. PMLR. JMLR.org, p. 1929-1938.

He, Zhenliang, Wangmeng Zuo, Meina Kan, Shiguang Shan et Xilin Chen (nov. 2019). « Attgan : Facial attribute editing by only changing what you want ». In : *IEEE Transactions on Image Processing* 28.11, p. 5464-5478. issn : 1057-7149. doi : 10.1109/tip.2019.2916751.

Holland, Sarah, Ahmed Hosny, Sarah Newman, Joshua Joseph et Kasia Chmielinski (2020). « The dataset nutrition label ». In : *Data Protection and Privacy, Volume 12 : Data Protection and Democracy* 12, p. 1. doi : 10.5040/9781509932771.ch-001.

Hutchinson, Ben et Margaret Mitchell (nov. 2019). « 50 years of test (un) fairness : Lessons for machine learning ». In : *Proceedings of the conference on fairness, accountability, and transparency*. ACM, p. 49-58. doi : 10.1145/3287560.3287600.

Ilvento, Christina (juin 2019). « Metric learning for individual fairness ». In : *arXiv preprint arXiv :1906.00250* 156. Sous la dir. d'Aaron Roth 0001, 2 :1-2 :11. doi : 10.4230/lipics.forc.2020.2.

Introna, Lucas D (juill. 1997). « Privacy and the computer : why we need privacy in the information society ». In : *Metaphilosophy* 28.3, p. 259-275. issn : 0026-1068.

Ismail Fawaz, Hassan, Germain Forestier, Jonathan Weber, Lhassane Idoumghar et Pierre-Alain Muller (mars 2019). « Deep learning for time series classification : a review ». In : *Data Mining and Knowledge Discovery* 33.4, p. 917-963. issn : 1573-756X. doi : 10.1007/s10618-019-00619-1.

Jagielski, Matthew, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi et Jonathan Ullman (déc. 2019). « Differentially private fair learning ». In : *International Conference on Machine Learning*. Sous la dir. de Kamalika Chaudhuri et Ruslan Salakhutdinov. T. 97. PMLR. PMLR, p. 3000-3008.

Jensen, Arthur R (1980). « Bias in mental testing ». In.

John, Philips George, Deepak Vijaykeerthy et Diptikalyan Saha (juin 2020). « Verifying individual fairness in machine learning models ». In : *Conference on Uncertainty in Artificial Intelligence*. Sous la dir. de Ryan P. Adams et Vibhav Gogate. PMLR. AUAI Press, p. 749-758.

Jones, Gareth P, James M Hickey, Pietro G Di Stefano, Charanpal Dhanjal, Laura C Stoddart et Vlasios Vasileiou (2020). « Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms ». In : *arXiv preprint arXiv :2010.03986* abs/2010.03986.

Jordan, Michael I et Tom M Mitchell (2015). « Machine learning : Trends, perspectives, and prospects ». In : *Science* 349.6245, p. 255-260.

Jourdan, Théo, Antoine Boutet et Carole Frindel (nov. 2018). « Toward privacy in IoT mobile devices for activity recognition ». In : *MobiQuitous*. Sous la dir. d'Henning Schulzrinne et Pan Li 0001. ACM, p. 155-165. doi : 10.1145/3286978.3287009.

Judge, Elizabeth F et Michael Pal (2021). « Voter Privacy and Big-Data Elections ». In : *Osgoode Hall LJ* 58, p. 1. issn : 1556-5068. doi : 10.2139/ssrn.3746632.

Justice, Minister of (1985). « Canadian Human Rights Act ». In : *RSC, 1985, c. H-6*.

Kallus, Nathan, Xiaojie Mao et Angela Zhou (jan. 2022). « Assessing algorithmic fairness with unobserved protected class using data combination ». In : *Management Science* 68.3. Sous la dir. de Mireille Hildebrandt, Carlos Castillo 0001, Elisa Celis, Salvatore Ruggieri, Linnet Taylor et Gabriela Zanfir-Fortuna, p. 1959-1981. doi : 10.1145/3351095.3373154.

Kamiran, Faisal et Toon Calders (fév. 2009). « Classifying without discriminating ». In : *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*. IEEE. IEEE, p. 1-6. doi : 10.1109/ic4.2009.4909197.

– (oct. 2012). « Data preprocessing techniques for classification without discrimination ». In : *Knowledge and Information Systems* 33.1, p. 1-33. issn : 0219-1377. doi : 10.1007/s10115-011-0463-8.

Kashid, Asmita, Vrushali Kulkarni, Ruhi Patankar et al. (2017). « Discrimination-aware data mining : a survey ». In : *International Journal of Data Science* 2.1, p. 70-84. issn : 2053-0811. doi : 10.1504/ijds.2017.082748.

Kearns, Michael, Seth Neel, Aaron Roth et Zhiwei Steven Wu (nov. 2017). « Preventing fairness gerrymandering : Auditing and learning for subgroup fairness ». In : *arXiv preprint arXiv :1711.05144* 80. Sous la dir. de Jennifer G. Dy et Andreas Krause 0001, p. 2569-2577.

Kilbertus, Niki, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing et Bernhard Schölkopf (juin 2017). « Avoiding discrimination through causal reasoning ». In : *Advances in neural information processing systems* 30. Sous la dir. d'Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan et Roman Garnett, p. 656-666.

Kim, Michael P, Amirata Ghorbani et James Zou (2019). « Multiaccuracy : Black-box post-processing for fairness in classification ». In : *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Sous la dir. de Vincent Conitzer, Gillian K. Hadfield et Shannon Vallor. ACM, p. 247-254. doi : 10.1145/3306618.3314287.

Kocaoglu, Murat, Christopher Snyder, Alexandros G Dimakis et Sriram Vishwanath (sept. 2017). « Causalgan : Learning causal implicit generative models with adversarial training ». In : *arXiv preprint arXiv :1709.02023*.

Krizhevsky, Alex, Ilya Sutskever et Geoffrey E Hinton (2012). « ImageNet Classification with Deep Convolutional Neural Networks ». In : *Advances in Neural Information Processing Systems*. Sous la dir. de F. Pereira, C.J. Burges, L. Bottou et K.Q. Weinberger. T. 25. Curran Associates, Inc. URL : https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kusner, Matt, Joshua Loftus, Chris Russell et Ricardo Silva (mars 2017). « Counterfactual fairness ». In : *Advances in neural information processing systems* 30. Sous la dir. d'Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan et Roman Garnett, p. 4066-4076.

Lahoti, Preethi, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang et Ed Chi (juin 2020). « Fairness without demographics through adversarially reweighted learning ». In : *Advances in neural information processing systems* 33. Sous la dir. d'Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan et Hsuan-Tien Lin, p. 728-740.

Lamy, Alex, Ziyuan Zhong, Aditya K Menon et Nakul Verma (2019). « Noise-tolerant fair classification ». In : *Advances in Neural Information Processing Systems* 32. Sous la dir. d'Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Edward A. Fox et Roman Garnett.

Le Quy, Tai, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang et Eirini Ntoutsi (mai 2022). « A survey on datasets for fairness-aware machine learning ». In : *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery* 12, e1452. issn : 1942-4787.

Leefeldt, Ed et Amy Danise (2021). *The Witness Against You : Your Car*.

Lewis, David (2013). *Counterfactuals*. John Wiley & Sons.

Li, Ninghui, Tiancheng Li et Suresh Venkatasubramanian (avr. 2006). « t-closeness : Privacy beyond k-anonymity and l-diversity ». In : *2007 IEEE 23rd international conference on data engineering*. IEEE. IEEE, p. 106-115. doi : 10.1109/icde.2007.367856.

Liaw, Richard, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez et Ion Stoica (juill. 2018). « Tune : A Research Platform for Distributed Model Selection and Training ». In : *arXiv preprint arXiv :1807.05118* abs/1807.05118.

Lipton, Zachary C, Alexandra Chouldechova et Julian McAuley (nov. 2018). « Does mitigating ML's impact disparity require treatment disparity ? » In : *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Sous la dir. de Samy Bengio, Hanna M.

Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi et Roman Garnett, p. 8136-8146.

Liu, Bo, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi et Zihuai Lin (2021). « When machine learning meets privacy : A survey and outlook ». In : *ACM Computing Surveys (CSUR)* 54.2, p. 1-36.

Loftus, Joshua R, Chris Russell, Matt J Kusner et Ricardo Silva (mai 2018). « Causal reasoning for algorithmic fairness ». In : *arXiv preprint arXiv :1805.05859* abs/1805.05859.

Lohia, Pranay K, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney et Ruchir Puri (mai 2019). « Bias mitigation post-processing for individual and group fairness ». In : *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE. IEEE, p. 2847-2851. doi : 10.1109/icassp.2019.8682620.

Luong, Binh Thanh, Salvatore Ruggieri et Franco Turini (2011). « k-NN as an implementation of situation testing for discrimination discovery and prevention ». In : *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. Sous la dir. de Chid Apt&eacute ; Joydeep Ghosh et Padhraic Smyth. ACM Press, p. 502-510. doi : 10.1145/2020408.2020488.

Lys, Igor (2019). *Data in Politics : an overview*.

Madras, David, Elliot Creager, Toniann Pitassi et Richard Zemel (fév. 2018). « Learning adversarially fair and transferable representations ». In : *arXiv preprint arXiv :1802.06309* 80. Sous la dir. de Jennifer G. Dy et Andreas Krause 0001, p. 3381-3390.

Malekzadeh, Mohammad, Richard G. Clegg, Andrea Cavallaro et Hamed Haddadi (avr. 2018). « Protecting Sensory Data Against Sensitive Inferences ». In : *W-P2DS'18*. Sous la dir. de Francisco Maia, Hugues Mercier et Andrey Brito. ACM, 2 :1-2 :6. doi : 10.1145/3195258.3195260.

– (avr. 2019). « Mobile sensor data anonymization ». In : *IoTDI*. Sous la dir. d'Olaf Landsiedel et Klara Nahrstedt, p. 49-58. doi : 10.1145/3302505.3310068.

Martin, Kirsten E (2019). « Designing ethical algorithms ». In : *MIS Quarterly Executive June* 18, p. 129-142. issn : 1540-1960. doi : 10.17705/2msqe.00012.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman et Aram Galstyan (juill. 2021). « A survey on bias and fairness in machine learning ». In : *ACM Computing Surveys (CSUR)* 54.6, p. 1-35. issn : 0360-0300. doi : 10.1145/3457607.

Microsoft (2022). *Our Approach to Responsible AI at Microsoft*. URL : https://www.microsoft.com/en-us/ai/our-approach (visité le 02/09/2022).

Ministre de l'innovation des sciences et de l'industrie (2020). « 11 : An Act to Enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to Make Consequential and Related Amendments to other Acts.(2020). 1st Reading November 17, 2020, 43rd Parliament ». In : *2nd Session. https ://parl.ca/DocumentViewer/en/43-2/bill/C-11/first-reading*. URL : https://parl.ca/DocumentViewer/en/43-2/bill/C-11/first-reading.

Mirza, Mehdi et Simon Osindero (nov. 2014). « Conditional generative adversarial nets ». In : *arXiv preprint arXiv :1411.1784* abs/1411.1784.

Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour et Kristian Lum (2018). « Prediction-based decisions and fairness : A catalogue of choices, assumptions, and definitions ». In : *arXiv preprint arXiv :1811.07867*.

Mitchell, Tom M. (1997). *Machine Learning*. New York : McGraw-Hill. isbn : 978-0-07-042807-2. doi : 10.1007/978-1-4613-2279-5.

Mohammadi, Kiarash, Aishwarya Sivaraman et Golnoosh Farnadi (juin 2022). « FETA : Fairness Enforced Verifying, Training, and Predicting Algorithms for Neural Networks ». In : *arXiv preprint arXiv :2206.00553* abs/2206.00553. doi : 10.48550/arxiv.2206.00553.

Moritz, Hardt (2013). *Fairness through Awareness*.

Mukherjee, Debarghya, Mikhail Yurochkin, Moulinath Banerjee et Yuekai Sun (juin 2020). « Two simple ways to learn individual fairness metrics from data ». In : *International Conference on Machine Learning*. T. 119. PMLR. PMLR, p. 7097-7107.

Muraski, Gregory (2021). *The Ugly Side of Beauty AI*. (Visité le 11/02/2022).

Narayanan, Arvind (2018a). « Translation tutorial : 21 fairness definitions and their politics ». In : *Proc. Conf. Fairness Accountability Transp., New York, USA*. T. 2. 3, p. 6-2.

– (2018b). « Translation tutorial : 21 fairness definitions and their politics ». In : *Proc. Conf. Fairness Accountability Transp., New York, USA*. T. 1170, p. 3.

Nations Unies (fév. 2008). *Déclaration universelle des droits de l'homme*. Département de l'information de l'ONU. isbn : 9789210553971. doi : 10.18356/2f12cb9b-fr.

O'Mahony, Niall, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan et Joseph Walsh (oct. 2019). « Deep learning vs. traditional computer vision ». In : *Science and information conference*. Sous la dir. de

Kohei Arai et Supriya Kapoor. T. 943. Springer. Springer International Publishing, p. 128-144. isbn : 9783030177942. doi : 10.1007/978-3-030-17795-9_10.

OHRC (2022). *Human Rights at Work 2008 - Third Edition*. (Visité le 15/06/2022).

Parlement Européen (2018). *Le règlement général sur la protection des données - RGPD*.

– (2021). *RÈGLEMENT DU PARLEMENT EUROPÉEN ET DU CONSEIL ÉTABLISSANT DES RÈGLES HARMONISÉES CONCERNANT L'INTELLIGENCE ARTIFICIELLE (LÉGISLATION SUR L'INTELLIGENCE ARTIFICIELLE) ET MODIFIANT CERTAINS ACTES LÉGISLATIFS DE L'UNION*. url : %7Bhttps://eur-lex.europa.eu/legal-content/FR/ALL/?uri=CELEX%5C%3A52021PC0206%7D.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai et Soumith Chintala (2019). « PyTorch : An Imperative Style, High-Performance Deep Learning Library ». In : *Advances in Neural Information Processing Systems 32*. Sous la dir. de H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox et R. Garnett. Vancouver, Canada : Curran Associates, Inc., p. 8024-8035.

Pearl, Judea (jan. 2009). « Causal inference in statistics : An overview ». In : *Statistics surveys* 3, p. 96-146. issn : 1935-7516.

Perera, Charith, Rajiv Ranjan, Lizhe Wang, Samee U Khan et Albert Y Zomaya (mai 2015). « Big data privacy in the internet of things era ». In : *IT Professional* 17.3, p. 32-39. issn : 1520-9202. doi : 10.1109/mitp.2015.34.

Pessach, Dana et Erez Shmueli (jan. 2020). « Algorithmic fairness ». In : *arXiv preprint arXiv :2001.09784* abs/2001.09784.

Pincus, FL (1996). « Discrimination comes in many forms : Individual ». In : *Institutional, and Structural„„The American Behavioral Scientist* 40.2.

Ping, Haoyue, Julia Stoyanovich et Bill Howe (2017). « Datasynthesizer : Privacy-preserving synthetic datasets ». In : *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM, p. 1-5. doi : 10.1145/3085504.3091117.

Popescu, Marius-Constantin, Valentina E Balas, Liliana Perescu-Popescu et Nikos Mastorakis (2009). « Multilayer perceptron and neural networks ». In : *WSEAS Transactions on Circuits and Systems* 8.7, p. 579-588.

PRIVACYINTERNATIONAL (2019). *Twitter may have used your personal data for ads without your permission. Time to fix AdTech !*

QUINLAN, J. Ross (1986). « Induction of decision trees ». In : *Machine learning* 1.1, p. 81-106.

RACHELS, James (nov. 1975). « Why privacy is important ». In : *Philosophy & Public Affairs*, p. 323-333.

RAVAL, Nisarg, Ashwin MACHANAVAJJHALA et Jerry PAN (jan. 2019). « Olympus : Sensor Privacy through Utility Aware Obfuscation ». In : *Proceedings on Privacy Enhancing Technologies* 2019.1, p. 5-25. ISSN : 2299-0984. DOI : 10.2478/popets-2019-0002.

RAWLS, John (juill. 2004). « A theory of justice ». In : *Ethics*. Routledge, p. 229-234. ISBN : 9780674042582. DOI : 10.2307/2693651.

REYES-ORTIZ, J. L. (2015). *Smartphone-based human activity recognition*. Springer. ISBN : 9783319142739. DOI : 10.1007/978-3-319-14274-6.

RIGAKI, Maria et Sebastian GARCIA (2020). « A survey of privacy attacks in machine learning ». In : *arXiv preprint arXiv :2007.07646*.

ROMEI, Andrea et Salvatore RUGGIERI (nov. 2014). « A multidisciplinary survey on discrimination analysis ». In : *The Knowledge Engineering Review* 29.5, p. 582-638. ISSN : 0269-8889.

ROSELLI, Drew, Jeanna MATTHEWS et Nisha TALAGALA (mai 2019). « Managing bias in AI ». In : *Companion Proceedings of The 2019 World Wide Web Conference*. Sous la dir. de Sihem AMER-YAHIA, Mohammad MAHDIAN, Ashish GOEL, Geert-Jan HOUBEN, Kristina LERMAN, Julian J. MCAULEY, Ricardo A. BAEZA-YATES et Leila ZIA. ACM, p. 539-544. DOI : 10.1145/3308560.3317590.

RUGGIERI, Salvatore (2014). « Using t-closeness anonymity to control for non-discrimination. » In : *Trans. Data Privacy* 7.2, p. 99-129.

RUGGIERI, Salvatore, Sara HAJIAN, Faisal KAMIRAN et Xiangliang ZHANG (2014). « Anti-discrimination analysis using privacy attack strategies ». In : *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Sous la dir. de Toon CALDERS, Floriana ESPOSITO, Eyke HÜLLERMEIER et Rosa MEO. T. 8725. Springer. Springer Berlin Heidelberg, p. 694-710. ISBN : 9783662448502. DOI : 10.1007/978-3-662-44851-9_44.

RUMELHART, David E, Geoffrey E HINTON et Ronald J WILLIAMS (1985). *Learning internal representations by error propagation*. Rapp. tech. California Univ San Diego La Jolla Inst for Cognitive Science, p. 399-421. DOI : 10.1016/b978-1-4832-1446-7.50035-2.

Ruoss, Anian, Mislav Balunović, Marc Fischer et Martin Vechev (fév. 2020). « Learning certified individually fair representations ». In : *arXiv preprint arXiv :2002.10312*. Sous la dir. d'Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan et Hsuan-Tien Lin.

Salazar, Ricardo, Felix Neutatz et Ziawasch Abedjan (mai 2021). « Automated feature engineering for algorithmic fairness ». In : *Proceedings of the VLDB Endowment* 14.9, p. 1694-1702. issn : 2150-8097. doi : 10.14778/3461535.3463474.

Salem, Ahmed, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz et Michael Backes (2018). « Ml-leaks : Model and data independent membership inference attacks and defenses on machine learning models ». In : *arXiv preprint arXiv :1806.01246*.

Samuel, Arthur L (jan. 1967). « Some studies in machine learning using the game of checkers. II—Recent progress ». In : *IBM Journal of research and development* 11.6, p. 601-617. issn : 0066-4138. doi : 10.1016/0066-4138(69)90004-4.

Shokri, Reza, Marco Stronati, Congzheng Song et Vitaly Shmatikov (mai 2017). « Membership inference attacks against machine learning models ». In : *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. IEEE, p. 3-18. doi : 10.1109/sp.2017.41.

Sidey-Gibbons, Jenni AM et Chris J Sidey-Gibbons (déc. 2019). « Machine learning in medicine : a practical introduction ». In : *BMC medical research methodology* 19.1, p. 1-18. issn : 1471-2288. doi : 10.1186/s12874-019-0681-4.

Silva, Selena et Martin Kenney (2018). « Algorithms, platforms, and ethnic bias : An integrative essay ». In : *Phylon (1960-)* 55.1 & 2, p. 9-37.

– (oct. 2019). « Algorithms, platforms, and ethnic bias ». In : *Communications of the ACM* 62.11, p. 37-39. issn : 0001-0782. doi : 10.1145/3318157.

Smith, Elaine (2021). *Is trading your privacy for lower insurance rates a good idea ?*

Song, Liwei, Reza Shokri et Prateek Mittal (mai 2019). « Membership inference attacks against adversarially robust deep learning models ». In : *2019 IEEE Security and Privacy Workshops (SPW)*. IEEE. IEEE, p. 50-56. doi : 10.1109/spw.2019.00021.

Soremekun, Ezekiel, Mike Papadakis, Maxime Cordy et Yves Le Traon (mai 2022). « Software Fairness : An Analysis and Survey ». In : *arXiv preprint arXiv :2205.08809* abs/2205.08809. doi : 10.48550/arxiv.2205.08809.

Sprager, S. et M. B. Juric (sept. 2015). « Inertial sensor-based gait recognition : a review ». In : *Sensors* 15.9, p. 22089-22127. issn : 1424-8220.

Srivastava, Akash, Lazar Valkov, Chris Russell, Michael U. Gutmann et Charles Sutton (2017). « VEEGAN : Reducing Mode Collapse in GANs using Implicit Variational Learning ». In : *Advances in Neural Information Processing Systems*. Sous la dir. d'I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett. T. 30. Curran Associates, Inc. URL : https://proceedings.neurips.cc/paper/2017/file/44a2e0804995faf8d2e3b084a1e2db1d-Paper.pdf.

Stilgoe, Jack (nov. 2018). « Machine learning, social learning and the governance of self-driving cars ». In : *Social studies of science* 48.1, p. 25-56. ISSN : 1556-5068. DOI : 10.2139/ssrn.2937316.

Sujay V., Lionel (2021). *Digital Assistants to Reach More Than 4 Billion Devices in 2017 as Google Set to Take a Lead, IHS Markit Says*. (Visité le 08/02/2022).

Suresh, Harini et John Guttag (oct. 2021). « A framework for understanding sources of harm throughout the machine learning life cycle ». In : *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, p. 1-9. DOI : 10.1145/3465416.3483305.

Suresh, Harini et John V Guttag (2019). « A framework for understanding unintended consequences of machine learning ». In : *arXiv preprint arXiv :1901.10002* abs/1901.10002.

Sweeney, Latanya (oct. 2002). « k-anonymity : A model for protecting privacy ». In : *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05, p. 557-570. ISSN : 0218-4885. DOI : 10.1142/s0218488502001648.

Tedesco, Salvatore, John Barton et Brendan O'Flynn (juin 2017). « A review of activity trackers for senior citizens : Research perspectives, commercial landscape and the role of the insurance industry ». In : *Sensors* 17.6, p. 1277. ISSN : 1424-8220.

Tian, Huan, Tianqing Zhu, Wei Liu et Wanlei Zhou (août 2022). « Image fairness in deep learning : problems, models, and challenges ». In : *Neural Computing and Applications* 34, p. 1-19. ISSN : 0941-0643. DOI : 10.1007/s00521-022-07136-1.

Toch, Eran, Yang Wang et Lorrie Faith Cranor (avr. 2012). « Personalization and privacy : a survey of privacy risks and remedies in personalization-based systems ». In : *User Modeling and User-Adapted Interaction* 22.1, p. 203-220. ISSN : 0924-1868. DOI : 10.1007/s11257-011-9110-z.

Tran, Cuong, Ferdinando Fioretto et Pascal Van Hentenryck (sept. 2021). « Differentially private and fair deep learning : A lagrangian dual approach ». In : *Proceedings of the AAAI Conference on Artificial Intelligence*. T. 35. 11. AAAI Press, p. 9932-9939.

Vavoulas, George, Charikleia Chatzaki, Thodoris Malliotakis, Matthew Pediaditis et Manolis Tsiknakis (2016). « The MobiAct Dataset : Recognition of Activities of Daily Living using

Smartphones. » In : *ICT4AgeingWell*. Sous la dir. de Carsten Röcker, Martina Ziefle, John O'Donoghue, Leszek A. Maciaszek et William Molloy. SCITEPRESS - Science, p. 143-151.

Veale, Michael et Irina Brass (avr. 2019). « Administration by algorithm ? Public management meets public sector machine learning ». In : *Public management meets public sector machine learning*.

Verma, Sahil et Julia Rubin (mai 2018). « Fairness Definitions Explained ». In : sous la dir. d'Yuriy Brun, Brittany Johnson et Alexandra Meliou, p. 1-7. doi : 10.1145/3194770.3194776.

Vincent, James (2018). *Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech*. (Visité le 11/02/2022).

– (2020). *AI camera operator repeatedly confuses bald head for soccer ball during live stream*. (Visité le 11/02/2022).

Wadsworth, Christina, Francesca Vera et Chris Piech (juin 2018). « Achieving Fairness through Adversarial Learning : an Application to Recidivism Prediction ». In : *arXiv preprint arXiv :1807.00199* abs/1807.00199.

Wang, Lipo (2005). *Support vector machines : theory and applications*. T. 177. Springer Science Business Media.

Wang, Serena, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta et Michael Jordan (fév. 2020). « Robust optimization for fairness with noisy protected groups ». In : *Advances in Neural Information Processing Systems* 33. Sous la dir. d'Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan et Hsuan-Tien Lin, p. 5190-5203.

Williamson, Robert et Aditya Menon (jan. 2019). « Fairness risk measures ». In : *International Conference on Machine Learning*. Sous la dir. de Kamalika Chaudhuri et Ruslan Salakhutdinov. T. 97. PMLR. PMLR, p. 6786-6797.

Woodworth, Blake, Suriya Gunasekar, Mesrob I Ohannessian et Nathan Srebro (2017). « Learning non-discriminatory predictors ». In : *Conference on Learning Theory*. Sous la dir. de Satyen Kale et Ohad Shamir. T. 65. PMLR. JMLR.org, p. 1920-1953.

Wu, Jiqing, Zhiwu Huang, Janine Thoma, Dinesh Acharya et Luc Van Gool (déc. 2018). « Wasserstein divergence for gans ». In : *Proceedings of the European Conference on Computer Vision (ECCV)*. Sous la dir. de Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu et Yair Weiss. T. 11209. Springer International Publishing, p. 653-668. isbn : 9783030012274. doi : 10.1007/978-3-030-01228-1_40.

Wu, Yongkai, Lu Zhang, Xintao Wu et Hanghang Tong (2019). « Pc-fairness : A unified framework for measuring causality-based fairness ». In : *Advances in neural information processing systems* 32.

Wyden, Ron (2022). *Algorithmic Accountability Act of 2022*. https ://www.wyden.senate.gov/imo/media/doc/2022 02-03 Algorithmic Accountability Act of 2022 One-pager.pdf.

Xu, Depeng, Yongkai Wu, Shuhan Yuan, Lu Zhang et Xintao Wu (août 2019). « Achieving causal fairness through generative adversarial networks ». In : *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. Sous la dir. de Sarit Kraus. International Joint Conferences on Artificial Intelligence Organization, p. 1452-1458. doi : 10.24963/ijcai.2019/201.

Xu, Depeng, Shuhan Yuan, Lu Zhang et Xintao Wu (déc. 2018). « FairGAN : Fairness-aware Generative Adversarial Networks ». In : *arXiv preprint arXiv :1805.11202*. Sous la dir. de Naoki Abe, Huan Liu 0001, Calton Pu, Xiaohua Hu, Nesreen Ahmed, Mu Qiao, Yang Song, Donald Kossmann, Bing Liu 0001, Kisung Lee, Jiliang Tang, Jingrui He et Jeffrey Saltz, p. 570-575. doi : 10.1109/bigdata.2018.8622525.

– (déc. 2019). « Fairgan+ : Achieving fair data generation and classification through generative adversarial nets ». In : *2019 IEEE International Conference on Big Data (Big Data)*. IEEE. IEEE, p. 1401-1406. doi : 10.1109/bigdata47090.2019.9006322.

Yudkowsky, Eliezer et al. (juill. 2008). « Artificial intelligence as a positive and negative factor in global risk ». In : *Global catastrophic risks* 1.303, p. 184.

Yurochkin, Mikhail, Amanda Bower et Yuekai Sun (juin 2019). « Training individually fair ML models with sensitive subspace robustness ». In : *arXiv preprint arXiv :1907.00020*.

Yurochkin, Mikhail et Yuekai Sun (juin 2020). « Sensei : Sensitive set invariance for enforcing individual fairness ». In : *arXiv preprint arXiv :2006.14168*.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez et Krishna P Gummadi (oct. 2017). « Fairness beyond disparate treatment & disparate impact : Learning classification without disparate mistreatment ». In : *Proceedings of the 26th international conference on world wide web*. Sous la dir. de Rick Barrett, Rick Cummings, Eugene Agichtein et Evgeniy Gabrilovich. ACM, p. 1171-1180. doi : 10.1145/3038912.3052660.

Zehlike, Meike, Ke Yang et Julia Stoyanovich (mars 2021). « Fairness in ranking : A survey ». In : *arXiv preprint arXiv :2103.14000* abs/2103.14000.

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi et Cynthia Dwork (2013). « Learning fair representations ». In : *International Conference on Machine Learning*. T. 28. JMLR.org, p. 325-333.

Zhang, Bo, Ruotong Yu, Haipei Sun, Yanying Li, Jun Xu et Hui Wang (jan. 2020). « Privacy for All : Demystify Vulnerability Disparity of Differential Privacy against Membership Inference Attack ». In : *arXiv preprint arXiv :2001.08855* abs/2001.08855.

Zhang, Brian Hu, Blake Lemoine et Margaret Mitchell (déc. 2018). « Mitigating unwanted biases with adversarial learning ». In : sous la dir. de Jason Furman, Gary E. Marchant, Huw Price et Francesca Rossi, p. 335-340. doi : 10.1145/3278721.3278779.

Zhang, Hantian, Xu Chu, Abolfazl Asudeh et Shamkant B Navathe (juin 2021). « Omnifair : A declarative system for model-agnostic group fairness in machine learning ». In : *Proceedings of the 2021 International Conference on Management of Data*. Sous la dir. de Guoliang Li 0001, Zhanhuai Li, Stratos Idreos et Divesh Srivastava. ACM, p. 2076-2088. doi : 10.1145/3448016.3452787.