UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Triennale in Informatica

TESI DI LAUREA

Fairness, Privacy, Ethics in sistemi di Machine Learning

RELATORE

Prof. Fabio Palomba

Università degli Studi di Salerno

CANDIDATO

Thomas De Palma

Matricola: 0512109541

Anno Accademico 2022-2023

Questa tesi è stata realizzata nel

"Arguing that you don't care about the right to privacy because you have nothing to hide is no different than saying you don't care about free speech because you have nothing to say."

Edward Snowden

Abstract

Al giorno d'oggi i computer sono diventati indispensabili per la vita di tutti i giorni. Ogni aspetto della nostra quotidianità é influenzato dai sistemi informatici, ad esempio i settori economici, sanitari, pubblicitari, etc. Particolare attenzione é stata suscitata dal rapido sviluppo dell'intelligenza artificiale e dalle implicazioni nella vita reale che conseguono dall'utilizzo di moduli intelligenti. Non sono rari, infatti, gli episodi in cui sono state registrate imparzialitá e discriminazioni nelle predizioni nei confronti di gruppi di persone, in base al sesso, etnia o religione. Conseguentemente al problema dell'equitá nei sistemi di machine learning, un forte sentimento di preoccupazione nella gestione delle informazioni private si é instaurato negli utenti. Gli episodi di attacchi da parte di malintenzionati a sistemi intelligenti sono sempre piú frequenti. Si rende quindi necessaria l'adozione di protocolli di sicurezza atti a garantire sia la privacy degli utenti che dei modelli di machine learning stessi. L'etica e la privacy sono due pilastri fondamentali su cui si basa la progettazione di un modulo di machine learning sicuro. L'obiettivo di questo lavoro di tesi è quello di fornire, tramite revisione sistematica della letteratura, una panormainca congiunta dei concetti di privacy e fairness, e delle relative implicazioni nello sviluppo di sistemi di machine learning. In particolare, viene posta attenzione sul riscontro in contesti reali in cui vengono applicati moduli di machine learning dove é possibile analizzare l'influenza reciproca tra privacy e fairness. Tra i principali risultati ottenuti si osserva che la misurazione dell'influenza reciproca tra i due concetti non é applicabile ad ogni contesto. In particolare, ogni problematica viene risolta attraverso soluzioni specifiche e spesso limitate al singolo caso di studio.

Indice

Elenco delle figure

Elenco delle tabelle

CAPITOLO 1

Introduzione

1.1 Motivazioni e Obiettivi

Lo sviluppo tecnologico ha decisamente introdotto numerose agevolazioni per il genere umano. La qualitá di vita é migliorata e attraverso le macchine siamo assistiti nella risoluzione di tante problematiche. L'analisi di grandi quantitá di dati tramite sistemi intelligenti ha aperto la societá a nuove prospettive di sviluppo ma anche a nuove criticitá da risolvere. Il crescente impiego di sistemi di machine learning in ambienti legati nostra quotidianità ha fatto sorgere preoccupazioni circa la tutela delle informazioni sensibili degli utenti. Analogamente alle problematiche relative alla tutela della privacy, sono state riscontrate ingiustizie involontarie ad opera di questi sistemi. Le questioni riguardanti i concetti di privacy e fairness alimentano considerevolmente la perdita di fiducia nei sistemi di intelligenza artificiale, generando danni morali e turbamenti psicologici. Sulla base di questa evidenza, gli obbiettivi di questo studio mirano a fornire maggiori informazioni circa le tematiche di equitá e privacy nei contesti in cui sono state riscontrate difficoltá, e laddove presenti vengono proposte tecniche di risoluzione a questi problemi. L'indagine condotta in questo studio mira ad ad osservare in quali ambiti di sviluppo i moduli di machine learning sono notoriamente impattati da problematiche che trattano in maniera congiunta

entrambi gli aspetti qualitativi, e le strategie messe in atto per mitigarle.

1.2 Struttura della tesi

Il lavoro di tesi viene suddiviso nei seguenti capitoli:

- **Capitolo 2: Background**, viene fornita una panoramica generale sui sistemi di Machine Learning, le fasi progettuali attraverso le quali un modulo é sottoposto,le varie tipologie di algoritmi utilizzati e i metodi di addestramento.
- **Capitolo 3: Stato dell'arte**, che presenta un'indagine individuale dei concetti di privacy e fairness nei moduli di Machine Learning in letteratura.
- **Capitolo 4: Metodologie di ricerca**, che illustra il metodo di ricerca adoperato per condurre l'analisi. Vengono fornite le domande di ricerca e le strategie adottate per ottenere risultati.
- **Capitolo 5: Analisi dei risultati**, che espone i risultati ottenuti dall'applicazione del metodo di ricerca, vengono fornite le risposte alle domande di ricerca.
- **Capitolo 6: Conclusioni**, viene effettuata una sintesi del lavoro svolto e vengono introdotti possibili lavori futuri.

CAPITOLO 2

Background

2.1 Intelligenza Artificiale e Machine Learning

Negli ultimi due decenni una silente rivoluzione si é svolta in ambito informatico. I computer, che dal loro avvento hanno digitalizzato e migliorato le nostre vite, sono i soggetti principali di questo cambiamento, non piú dei semplici calcolatori numerici come un tempo, bensí gli strumenti che hanno dato via alla cosidetta terza rivoluzione industriale e alla societá dell'informazione. Al giorno d'oggi esistono sempre piú programmi che "imparano" e adattano il loro comportamento di volta in volta per eseguire problemi sempre piú complessi. Ogni tipo di informazione, partendo da semplici numeri a immagini, video e audio, é processata e trasferita digitalmente sotto forma di dati. L'elevato quantitativo di informazione prodotta ha suscitato l'interesse nell'analisi dei dati e nel machine learning. L'ideologia alla base di queste discipline risiede nella credenza che dietro complesse e voluminose quantitá di dati esista un modello che ne semplifichi lo studio.Il machine learning (ML) dunque non é una semplice applicazione per estrarre informazioni dai dati, l'apprendimento é un requisito fondamentale e caratterizzante per l'intelligenza. Un sistema intelligente deve potersi adattare all'ambiente, deve poter imparare a non ripetere gli stessi errori bensí a replicare un successo [1]. Precedentemente a questa rivoluzione digitale, i

ricercatori sostenevano che per ottenere l'intelligenza artificiale fosse necessario un nuovo paradigma di ragionamento, nuovi modelli di computazione e nuovi algoritmi.Grazie al successo odierno dei sistemi intelligenti, sappiamo che tutto ció di cui abbiamo bisogno risiede nell'utilizzo di una voluminosa quantitá di dati di esempio e una potenza computazionale sufficiente per poter applicare algoritmi su tali dati [1]. Seppur l'intelligenza artificiale sia diventata una tematica fortemente discussa soltanto negli ultimi tempi, la teorizzazione di un sistema intelligente pone le fondamenta giá nella metá del secolo scorso. Il termine machine learning fu coniato nel 1959 da Arthur Lee Samuel, un impiegato della IBM, che lo definisce come un campo di studi che attribuisce ai computer la facoltá di apprendere senza che essi siano esplicitamente programmati [2]. Uno dei primi esperimenti per riprodurre una macchina intelligente avvenne nel 1960. La Raytheon Company sviluppó un sistema, Cybertron, per analizzare i segnali dei sonar e distinguere se i segnali rimbalzavano contro un sottomarino o contro un mammifero. Anche un normale computer poteva svolgere una compito simile, ma soltanto attraverso una complicata programmazione che fornisse al calcolatore precise istruzioni come farlo.Il responsabile dello sviluppo della Raytheon preferí perció emulare il processo di apprendimento umano, attraverso tentativi ed errori e associazioni tra esperienze passate e circostanze attuali. Cybertron prendeva in input migliaia di varietá di suoni ed era equipaggiato con un "goof button" che un operatore umano premeva ogni volta che sbagliava a computare il risultato. [3, 4].

Ad oggi il machine learning è utilizzato in numerose applicazioni, come il riconoscimento delle immagini, l'elaborazione del linguaggio naturale, sistemi di rilevamento delle frodi, veicoli autonomi, previsioni climatiche e strategie di marketing. Negli ultimi anni, si sta prendendo sempre piú consapevolezza del fatto che l'uso combinato di discipline come la statistica, la teoria dell'informazione e il machine learning, stia portando alla creazione di una scienza sempre più robusta e affidabile, basata su solide fondamenta matematiche, e strumentazioni avanzate e potenti. Il machine learning è considerato come un sottoinsieme, oltre che un requisito fondamentale, dell'intelligenza artificiale che combina i processi logici associati all'apprendimento umano attraverso l'utilizzo di algoritmi computazionali. Questi algoritmi sfruttano grandi insiemi di dati (dataset) affinché sia possibile istruire la macchina a riconosce-

re dei pattern e a compiere decisioni autonome. Una branca del machine learning é strettamente collegata alla disciplina della statistica computazionale che utilizza big-data per ottenere complesse predizioni e sistemi decisionali. Attraverso delle opportune modifiche e ripetizioni di un algoritmo, la macchina è in grado di prevedere un output dato un insieme di input. Gli output sono poi messi a confronto con dei risultati noti come "veri" per poter giudicare la precisione delle previsioni, questo meccanismo viene iterato piú volte fino ad ottenere delle predizioni quanto piú realistiche possibili [5].

Al fine di progettare un modulo di machine learning in maniera corretta esistono diversi modelli di sviluppo e passaggi ben strutturati da seguire. Un esempio é l'approccio proposto da De Cristofaro [6] che prevede i seguenti passaggi:

1. **Allenamento**, una volta che i dati sono stati collezionati e pre-processati viene scelto un modello di machine learning da allenare. Un modello di machine learning puó essere astratto ad una funzione parametrica $h_\theta(x)$ che prende un input x e un vettore parametrico θ. L'input x é spesso rappresentato come un vettore di valori chiamati features. Durante questa fase il modello analizza i dati di allenamento per trovare i valori dei parametri θ. In altre parole, durante il training, un algoritmo di ML ha come obbiettivo l'apprendimento di informazioni rispetto ad una determinata task da svolgere.

2. **Inferenza/Testing**, la performance del modello viene verificata su un dataset di prova che deve essere indipendente dal dataset utilizzato per l'allenamento, affinché si possano valutare e misurare le capacitá dell'algoritmo di effettuare previsioni. Ci si rende certi che l'algoritmo abbia compreso il proprio compito. In maniera formale i valori dei parametri θ sono fissati e il modello computa la funzione $h_\theta(x)$ con nuovi input x. La predizione del modello puó assumere forme diverse, ad esempio per i problemi di classificazione la piú comune è un vettore che assegna ad ogni classe del problema una probabilitá che caratterizza l'appartenenza di un dato input a una determinata classe.

I problemi risolti attraverso il machine learning sono comunemente divisi in tre tipologie, caratterizzate in base alla struttura dei dati analizzati e dal corrispondente algoritmo di apprendimento utilizzato [6, 7]:

- **Apprendimento supervisionato – Supervised Learning,** consiste in una serie di tecniche che inducono una associazione tra i dati in input e output basandosi su esempi che sono giá stati messi in relazione tra loro. Se l'insieme dei risultati in output definisce una categoria, allora si tratta di un problema di classificazione, ad esempio il riconoscimento di oggetti tramite immagini. Si parla invece di problemi di regressione se l'obbiettivo è la previsione di una variabile numerica, come il prezzo di un oggetto in base alle sue caratteristiche.

- **Apprendimento non supervisionato – Unsupervised Learning**, consiste in metodi di allenamento che forniscono in input dati non categorizzati. L'algoritmo dovrá identificare le relazioni tra le informazioni da solo con l'obbiettivo di trovare pattern nascosti. Questa tipologia di apprendimento viene utilizzata per raggruppare dati simili in cluster o riconoscere eventuali anomalie.

- **Apprendimento rinforzato – Reinforcement Learning**, analogamente all'apprendimento non supervisionato non vengono forniti dataset giá categorizzati. L'algoritmo viene addestrato attraverso ricompense o "penalitá" per l'esecuzione di determinate task, in questo modo impara e si adatta sulla base di tentativi ed errori basandosi sul feedback ricevuto. Questa tipologia di allenamento trova largo impiego in applicazioni nel settore della robotica e della guida autonoma.

Un'ulteriore distinzione è data dall'architettura dei modelli di machine learning, esistono infatti due configurazioni principali [8]:

- **Apprendimento centralizzato - Centralized learning**, i dataset di addestramento sono conservati in una singola entitá, una macchina o un data center fungono da server centrale per coordinare ed eseguire training sui modelli di ML. In questa architettura il server ha accesso ai dati di ogni dispositivo connesso permettendo lo sviluppo di modelli sempre piú efficienti e precisi. Tuttavia, poiché un operatore centrale ha accesso diretto a tutte le informazioni sensibili, sono state sollevate preoccupazioni riguardanti la privacy. In questo modo, infatti , si è piú esposti ad accessi non autorizzati e data breach [8]. Ulteriori svantaggi sono dovuti dall'enorme quantitá di spazio di archiviazione necessaria per conservare i dataset.

- **Apprendimento distribuito - Distributed learning**, in questo scenario il processo di training viene decentralizzato su piú dispositivi o macchine. Ogni nodo opera su una frazione del dataset e periodicamente scambia informazioni con le altre macchine per aggiornare i parametri del modello. Lo studio condotto da Liu et. al propone alcune variazioni dell'apprendimento distribuito [8]:
  - Collaborative learning;
  - Federated learning;
  - Split learning.

Un'altro approccio per istanziare un modello di machine learning é attraverso le reti neurali, dove l'apprendimento puó essere sia supervisionato che non-supervisionato.Una rete neurale é composta da tanti processori , chiamati neuroni, interconnessi tra loro, spesso organizzati in livelli a cui é assegnato un peso. Ogni nodo effettua delle semplici computazioni sui dati in input e ad ogni livello vengono estratte informazioni sempre piú complesse fino a produrre l'output nel livello finale. Durante il training del modello il valore iniziale dei pesi é settato randomicamente e ogni qualvolta i dati passano attraverso gli strati della rete il valore viene bilanciato. Questo processo di bilanciamento si ripete finché attraverso i dati di addestramento con le stesse etichette si riescono a produrre output simili. Questa metodologia é meglio nota come deep learning, dove la parola "deep" rappresenta la profonditá degli strati attraverso i quali i dati sono processati [6]. L'approccio piú comune all'implementazione di moduli di machine learning é attraverso gli algoritmi. Alcuni esempi di algoritmi che risolvono task di supervised learning comprendono [6]:

- Alberi decisionali
- Regressione lineare
- Algoritmi di clustering
- Alberi decisionali casuali (random forest)
- Macchine a vettori di supporto

Una parte di questi algoritmi vengono definiti come "classificatori", provano cioè a categorizzare i dati in classi e categorie diverse in base agli input forniti. Un esempio di problema di classificazione ben noto in letteratura é quello della classificazione delle e-mail sulla base del loro contenuto. L'obbiettivo principale consiste nella differenziazione delle e-mail tra Spam e Non Spam sulla base dell'attendibilitá di parole, frasi e pattern presenti nel corpo di una e-mail. Durante l'addestramento vengono dati in input al classificatore dei dataset contenenti e-mail giá etichettate affinché possa riconoscere in futuro quale di queste sono Spam, ovvero messaggi non richiesti inviati in massa per fini pubblicitari o di phishing. Questa tipologia di algoritmo è utile per filtrare automaticamente la posta in arrivo e prevenire minacce e attivitá fraudolente. Lo studio dei classificatori è una tematica ampiamente discussa in letteratura, é molto complesso generare classificatori che non soffrano di alcuni problemi noti, ad esempio la stretta dipendenza dal dataset di partenza o il fenomeno del *garbage-in garbage-out*, dove la presenza di bias nei dataset di addestramento tende a riflettersi nelle previsioni del modulo stesso [9]. Nell'ambito dell'etica software, un esempio pratico puó essere dato da un classificatore addestrato con un dataset che presenta bias su determinati attributi sensibili, quali razza, sesso o religione. Tale classificatore puó effettuare predizioni imparziali, prediligendo o svavoreggiando una classe di individui rispetto ad un altra [9].

Difatti per garantire l'efficacia di un modulo machine learning é cruciale rispettare alcuni vincoli qualitativi che nella disciplina dei Requirements Engineering sono meglio noti come requisiti non funzionali. I requisiti non funzionali di un'applicazione che incorpora l'apprendimento automatico possono essere diversi da quelli di un sistema software tradizionale, tra i piú determinanti abbiamo performance, fairness, privacy e trasparenza [10].

CAPITOLO 3

Stato dell'arte

3.1 Machine Learning Privacy

L'utilizzo sempre piú frequente di moduli di machine learning all'interno di sistemi software ha portato l'attenzione dei ricercatori e degli ingegneri ad adottare tecniche e metodologie piú efficaci finalizzate a preservare la privacy degli utilizzatori. Lo scandalo di Facebook sui dati sensibili del 2018 ha determinato negli utenti una consapevolezza sempre maggiore del concetto di privacy [11, 8]. Prima di fornire una panoramica degli studi attualmente presenti in letteratura è opportuno definire formalmente il concetto di privacy. La privacy è il diritto di ogni individuo di poter controllare e fornire le autorizzazioni relative alle proprie informazioni personali affinché possano essere raccolte, archiviate, elaborate e distribuite. Questa definizione enfatizza l'importanza del controllo sulle informazioni personali riconoscendo la privacy come un diritto fondamentale per la libertá e diginitá individuale. È importante notare come il concetto di privacy nella disciplina del machine learning, venga declinato in maniera differente rispetto alla concezione tradizionale di privacy dei dati, in quanto le tecniche di machine learning possono essere sia avverse che di aiuto alla protezione delle informazioni individuali. Le metriche attualmente piú utilizzate sono basate sui concetti di entropia e probabilitá [11]. In letteratura sono presenti

numerose pubblicazioni [12, 13, 14] che dimostrano come i modelli di machine learning e i dataset di allenamento possano essere vulnerabili e vittima di attacchi che mirano alla privacy, portando ad una perdita di informazioni sensibili. Ad esempio attraverso il Deep Learning, una delle tecniche di ML che sfrutta reti neurali a piú livelli, si possono estrarre informazioni sensibili dai dati, anche se non sono stati esplicitamente definiti. Questi sono noti come "inference attacks" e possono rivelare informazioni che gli individui non intendevano divulgare. Quando un utente condivide un'immagine sui social network, un modello di deep learning, utilizzato da un malintenzionato, puó rivelare informazioni sensibili sull'utente, la posizione o anche il nome, per proteggersi è necessario pre-processare i dati condivisi attraverso tecniche di offuscamento e perturbazione delle informazioni. Oltre questi esempi noti di vulnerabilitá connesse alla privacy, altri studi di ricerca hanno dimostrato che si puó utilizzare il machine learning come tecnica per garantire la privacy, ad esempio Jun Yu et al. hanno sviluppato uno strumento per ottenere dei suggerimenti automatici delle impostazioni sulla privacy per la condivisione delle immagini [15]. I lavori esistenti in materia di privacy categorizzano i sistemi di machine learning in tre gruppi principali [8]:

- Sistemi di ML privati - Private Machine Learning, in questa classe l'obbiettivo principale consiste nel rendere il modello di machine learning e i dataset sicuri, poiché le minacce alla privacy possono avvenire in qualsiasi fase del processamento dei dati, ad esempio durante l'apprendimento, la pubblicazione o la predizione. La maggior parte della ricerca in questo ambito si basa sull'utilizzo della privacy differenziale e algoritmi di deep learning.

- Sistemi di ML per il migliorare la protezione dei dati, in questo caso gli algoritmi sono utilizzati come strumento per ottimizzare e garantire la privacy.

- Sistemi di ML finalizzati ad identificare e correggere le violazioni della privacy.

È importante notare come uno stesso modello di machine learning possa appartenere a piú categorie ed essere al contempo sia uno strumento di attacco che di protezione, rendendo l'analisi del problema ancora piú complicata. Il concetto di privacy é strettamente legato alla sicurezza di un sistema, se un modello é vulnerabile

come si puó garantire la privacy di chi lo utilizza? La sicurezza di ogni sistema è misurata rispetto alle capacitá di potersi difendere da un presunto attacco, viene perció introdotto il threat model [6]. Il threat model è un approccio strutturato per identificare e valutare le possibili minacce ad un sistema o ad una organizzazione. Per poter essere elaborato richiede un'analisi delle risorse, delle persone o dei sistemi che possono rappresentare una minaccia per tali risorse e le potenziali vulnerabilitá e impatti di un'attacco informatico. Attraverso l'implementazione di un threat model si possono quindi meglio comprendere le debolezze di un sistema e adottare misure per renderlo piú sicuro [7].

3.1.1 Private Machine Learning

In questo paragrafo verrá effettuata una tassonomia del threat model e verranno discusse le sfide e le soluzioni attuali che riguardano la tutela della privacy nei moduli di machine learning. Per garantire la privacy in un sistema di machine learning é necessario renderlo sicuro da attacchi, segue un'analisi quindi dal punto di vista di un malintenzionato. Come giá accennato precedentemente, gli obbiettivi di un attaccante sono i dataset di allenamento, che possono contenere dati sensibili, o l'algoritmo stesso. Ad esempio qualche istituzione finanziaria puó detenere dei modelli di machine learning sofisticati che sono in grado di effettuare predizioni accurate sull'andamento delle azioni, rendendo tale algoritmo un target appetibile per un criminale informatico [8]. Determinati gli obbiettivi avversari, segue un'indagine su quali sono i livelli di accesso e la conoscenza di un attaccante ed infine le metodologie di violazione informatica. La percezione di un modello di machine learning da parte di un hacker puó essere di due tipi [6, 8]:

- **White-Box**, l'attaccante ha delle informazioni sulla struttura dell'algoritmo o sul dataset;
- **Black-Box**, l'attaccante non ha alcuna informazione a priori, puó effettuare una analisi sul modello attraverso la sottomissione di input accuratamente selezionati e studiarne i risultati.

Un'altra variabile da tener presente è il momento in cui un attacco puó avvenire, in tal senso in letteratura si considerano due fasi ben specifiche per trattare questo tipo di aspetto [7]:

- **Fase di Inferenza**, definiti come attacchi esplorativi, non interferiscono con il modello di machine learning, piuttosto mirano ad ottenerne le caratteristiche. Gli attacchi durante questa fase possono presupporre sia una percezione white-box che black-box del modello.

- **Fase di training**,l'obbiettivo principale degli attacchi è influenzare o corrompere il modello stesso. Esistono due strategie principali per alterare il modello di machine learning. Nella prima viene corrotto il dataset di allenamento inserendovi input manipolati (injection), nella seconda vengono corrotti i dati di training in modo diretto. Nel caso del reinforcement learning l'attaccante puó modificare l'ambiente simulato all'interno del quale il modello si sta allenando.

Le attuali modalitá di attacco ad un modello di machine learning possono essere divise in quattro gruppi principali [8]:

- **Model Extraction Attacks**, questa categoria di attacchi presuppone una conoscenza black-box del modello. L'obiettivo dell'attaccante é ottenere un duplicato del modello di machine learning. Per ottenere i parametri interni o l'architettura puó interagire con le funzionalitá offerte da esso oppure attraverso uno studio degli output. L'hacker può utilizzare il modello estratto per creare un modello sostitutivo in grado di imitare il comportamento del modello originale oppure utilizzare le conoscenze acquisite dal modello per rivelare informazioni riservate o utilizzate nel processo decisionale e di addestramento del modello.

- **Feature Estimation Attacks**, sono mirati ad ottenere alcune proprietá statistiche del dataset di allenamento. Possono essere implementati attraverso il model inversion o shadow model.

- **Membership Inference Attacks**, in questa tipologia di attacchi l'obiettivo consiste nell'identificare l'appartenenza di un determinato record al dataset di allenamento. Shokri et al. [16] introducono il "black-box membership inference"

dove attraverso una tecnica di shadow training viene imitato il comportamento del modello vittima. Lo shadow-model viene allenato per riconoscere le differenze nelle predizioni del modello vittima, in particolare per classificare quali input sono stati utilizzati e quali no. Nella ricerca in questione viene sottolineato che le cause principali di una vulnerabilitá ad attacchi di membership inference sono dovute alla struttura utilizzata e al tipo di modello [16].

- **Model Memorization Attacks**, in questa categoria l'avversario cerca di ricostruire il dataset usato per l'addestramento di un modello di machine learning attraverso l'analisi del modello stesso.

Dopo aver illustrato tutte le tecniche di attacco, vengono successivamente riportati alcuni tra i piú importanti metodi di difesa che possono essere sfruttati per implementare un modulo di machine learning.

- **Crittografia**, puó essere usata per proteggere la confidenzialitá dei dati. La tecnica piú comune è l'homomorphic encryption (FHE), permette ad una macchina di effettuare computazioni su dati crittografati senza doverli decifrare preservandone la confidenzialitá. Si è dimostrato che applicare questa tecnica di crittografia al dataset di allenamento puó rallentare la computazione anche sui piú semplici problemi di classificazione [8]. La riduzione della complessitá computazionale rappresenta una delle maggiori sfide per la ricerca, soprattutto per quanto riguarda gli algoritmi di deep learning, che hanno tempi di calcolo lenti anche sfruttando dati non criptati. Un'altra tecnica diffusa è la secure-multiparty computation (SMC) che permette a due o piú macchine la computazione di una funzione in modo congiunto senza esporre i dettagli dei componenti utilizzati e degli input. Nella disciplina del machine learning, i metodi basati su SMC garantiscono la privacy sia del modello che dei dati [6].

- **Privacy Differenziale – Differential Privacy (DP)**, una tecnica utilizzata per preservare la privacy attraverso la condivisione dei dati senza rivelare peró informazioni sugli individui. Dal punto di vista di un attaccante dovrebbe essere indistinguibile stabilire se un certo dato appartiene al dataset di input oppure no. Attraverso la privacy differenziale è possibile condurre un'analisi sui dati in modo aggregato, preservando cosí le informazioni sensibili dei

singoli individui. Una delle modalitá piú comuni per implementare questo concetto, è attraverso l'offuscamento dei dati. Questo meccanismo mira a ridurre la precisione sui dati o del modello attraverso l'aggiunta di rumore. Uno dei metodi piú utilizzati nel campo del deep learning è il fast gradient sign method (FGSM), che aggiunge una piccola perturbazione ai dati in input per poter valutare, attraverso un'equazione lineare, l'impatto che tale perturbazione ha sui dati [11].

3.2 Fairness in Machine Learning

L'utilizzo sempre piú crescente dell'Intelligenza Artificiale per prendere decisioni importanti, rende necessaria la precauzione affinché tali decisioni non riflettano un comportamento discriminatorio verso alcuni gruppi o popolazioni [17]. Oltre i possibili utilizzi errati della tecnologia, è in aumento la preoccupazione che decisioni prese sulla base di algoritmi e dati non siano neutrali e possono amplificare disuguaglianze strutturali presenti e passate. Allo stato attuale esistono celebri esempi di spiacevoli inconvenienti che sono scaturiti da problemi di bias all'interno di sistemi software, di seguito vengono riportati i settori in cui si sono registrate discriminazioni:

- **Crime**, COMPAS è un software utilizzato negli Stati Uniti che misura la tendenza di una persona con precedenti penali a commettere di nuovo un crimine. I giudici sfruttano COMPAS per decidere se rilasciare una persona o prolungarne il periodo di detenzione. Un' indagine sul software ha fatto emergere che erano presenti dei pregiudizi sulle predizioni, in particolare verso Afro-Americani. COMPAS ha piú probabilitá di prevedere che un criminale Afro-Americano sia recidivo rispetto ad uno Caucasico [17].

- **Hiring Software nel campo STEM** (Science, Technology, Engineering, Math), consistono in software per la pubblicazione di annunci lavorativi nell'ambito del recruiting. Inizialmente gli algoritmi erano stati concepiti per poter essere gender – neutral, in questo modo le offerte di lavoro dovevano apparire in egual misura sia a donne che ad uomini. L'algoritmo in questione è stato testato in 191 nazioni del mondo e Anja Lambrecht e Catherine Tucker [18],

attraverso un'indagine statistica dimostrano che gli annunci avevano un tasso di comparsa del 20% in piú agli uomini rispetto alle donne. Questa differenza risulta particolarmente evidente nel range di etá compresa tra i 25-54 anni [18].

- **Facial Recognition Software**, consistono in algoritmi di classificazione che identificano oggetti o persone all'interno di immagini. Nel 2015 il famoso algoritmo di Google Foto per il riconoscimento facciale ha identificato persone di colore come "gorilla". Un problema analogo si è riscontrato con il software di riconoscimento facciale di Amazon, nel 2018 infatti i ricercatori hanno notato la presenza pregiudizi significativi contro persone di colore in particolare verso donne con tonalitá della pelle piú scure. Altro progetto degno di nota é "Gender Shades" condotto dal MIT Media Lab e Microsoft, i cui i risultati dimostrano come algoritmi di analisi facciale sviluppati da aziende del calibro di IBM, Microsoft e Megvii sono portati a valutare erroneamente le donne di pelle scura a causa di configurazioni errate sulle feature di addestramento, quali il sesso dell'individuo [19].

- **Autonomous Systems**, la possibilitá di un bias algoritmico suscita particolari preoccupazioni in sistemi di guida autonomi che coinvolgono esseri umani nei processi. Infatti man mano che tali sistemi diventano sempre piú complessi diventa altrettanto difficile comprendere come essi arrivano alle loro decisioni. Nel 2018 uno studio condotto dal Georgia Institute of Technology ha dimostrato come le auto con guida autonoma avevano piú probabilitá di colpire un pedone di colore rispetto ad un pedone di altra etnia. Lo studio è stato effettuato in un ambiente simulato all'interno del quale erano presenti numerosi scenari e pedoni di razze diverse [20].

- **Healthcare**, i sistemi sanitari si basano su algoritmi di previsione commerciali per identificare e aiutare i pazienti con esigenze di salute complesse. Un algoritmo sfruttato nel campo della sanitá negli Stati Uniti viene usato per determinare quali pazienti avranno bisogno di piú cure mediche e trattamenti speciali. L'algoritmo prende in considerazione una serie di parametri quali etá, sesso, condizioni di salute e razza. La ricerca condotta da Z. Obermeyer [21] ha dimostrato che usando la razza come parametro si puó incorrere in problemi di

Fairness, l'algoritmo ha dato prioritá a pazienti bianchi piuttosto che a pazienti di altre etnie. L'algoritmo tiene in considerazione anche i costi spesi in cure mediche , e questi ultimi essendo nettamente inferiori per le persone di colore, ha portatato all'erronea interpretazione che le persone di colore fossero piú in salute dei bianchi [21].

Con il termine biased-algorithm ci si riferisce dunque al fenomeno attraverso il quale modelli di machine learning producono predizioni e risultati ingiusti e discriminanti verso individui o gruppi di persone. Come espresso in letteratura il livello di fairness è strettamente collegato al concetto di bias, letteralmente pregiudizio. Le piú comuni cause di bias algoritmico includono:

- **Biased Data**, il pregiudizio puó essere introdotto nei modelli di machine learning giá a partire dai dataset sui quali il modello è stato allenato. Se il dataset contiene pattern discriminatori è molto probabile che tali pattern vadano ad inficiare sulle predizioni ottenute dall'algoritmo.

- **Limited Features**, i modelli di machine learning possono effettuare predizioni discriminatorie se sono stati implementati utilizzando un insieme limitato di funzionalitá che non aderiscono totalmente alla complessitá del problema da risolvere.

- **Design Choices**, spesso gli algoritmi possono generare comportamenti discriminatori se sono state fatte delle scelte di design inopportune, anche se i dati in sé utilizzati sono etici ed equi.

I risultati di questi biased-algorithms possono essere reintrodotti nel mondo reale influenzando ulteriori dataset che verranno utilizzati per l'allenamento di algoritmi futuri [17]. Conseguentemente a questa problematica i ricercatori, oltre a sviluppare migliori metodi di machine learning, si stanno approcciando alla software fairness concependo nuovi strumenti di testing per identificare e misurare i livelli di discriminazione. Nella seguente sezione verranno elencate le metriche e le definizioni del concetto di fairness affinché possa essere trattato come un argomento di rilevante importanza nello sviluppo di un software.

3.2.1 Definizioni e Metriche di Fairness

Implementare un algoritmo eticamente corretto, significa definire formalmente cosa si intende per Fairness, e cercare in conseguenza di misurare come ciò viene implementato. Fairness è un concetto che viene generalmente compreso trattando le persone in egual modo, senza pregiudizi o discriminazioni, affinché ad ognuno vengano date le stesse opportunitá.
Prima ancora della nascita dell'informatica, la filosofia e la psicologia hanno provato a dare una definizione di equitá, non riuscendo comunque a fornire una definizione universale [22]. Aristotele fu uno dei primi a dare un contributo marcante ai dibattiti moderni sull'equitá, nel suo lavoro "Etica Nicomachea" introduce infatti il concetto di *epieíkeia*, un principio che concepisce una forma di giustizia che va al di lá della legge scritta. Le decisioni dovevano essere prese sulla base di particolari circostanze e contesto riconoscendo che ogni individuo ha bisogni e capacitá differenti. L'equitá rappresenta una condizione necessaria per l'ottenimento della giustizia. [23]
È importante notare come spesso l'idea di Fairness possa essere declinata in significati differenti a seconda della persona a cui lo si chiede, queste variazioni di interpretazioni e definizioni possono essere dovute a ció che è piú o meno giusto a seconda delle influenze socio-culturali e dei contesti storici dei soggetti coinvolti. Gli esseri umani infatti, a differenza delle macchine che prendono decisioni sulla base di una sequenza logica di operazioni, sono condizionati da tanti fattori nel momento in cui bisogna effettuare un giudizio. Tali giudizi risultano inevitabilmente associati al modo di ragionare umano, quindi influenzati da emozioni, dalle esperienze vissute, da bias cognitivi e dal contesto sociale in cui l'uomo vive, sono dunque soggettivi.
Essendo legata alla natura soggettiva umana, in letteratura vengono introdotte oltre 20 differenti definizioni di Fairness nelle discipline legate all'Intelligenza artificiale e non esiste ancora un accordo chiaro su quale utilizzare in ogni situazione, ogni problema va analizzato caso per caso [24]. Esistono differenti modi di categorizzare le definizioni di fairness, una buon approccio consiste nel ragionare in termini di definizioni di gruppo o individuali. Di seguito vengono riportati vantaggi e svantaggi [25].
Nel contesto dell'intelligenza artificiale, le definizioni statistiche di Fairness sono

usate per misurare ed assicurarsi che un sistema AI tratti differenti individui o gruppi in modo equo. Questa famiglia di definizioni fissa un piccolo numero di gruppi demografici protetti, ovvero gruppi di persone che storicamente sono stati svantaggiati o discriminati sulla base di certe caratteristiche, ed effettua una stima approssimativa, attraverso una misurazione statistica, dell'equitá tra i gruppi stessi. Fondamentalmente vengono analizzati i risultati prodotti da sistemi di intelligenza artificiale e messi a confronto tra diversi gruppi demografici. Tuttavia, le definizioni statistiche di equità non danno di per sé garanzie significative per gli individui o i sottogruppi strutturati dei gruppi demografici protetti, bensí costituiscono delle garanzie per la media degli individui di un sottogruppo. La maggior parte delle misurazioni statistiche si basano sull'utilizzo di alcune metriche che possono essere meglio comprese sfruttando una matrice di confusione, dove le righe e le colonne rappresentano rispettivamente le istanze di predizione e le istanze attuali di una classe. Le metriche utilizzate sono le seguenti [24]:

- **Reale positivo - True positive (TP)**: una situazione in cui sia l'istanza di predizione che l'istanza attuale di una classe sono positive.
- **Falso positivo - False positive (FP)**: una situazione in cui il risultato predetto è positivo ma dovrebbe essere negativo.
- **Falso negativo - False negative (FN)**: una situazione in cui il risultato predetto è negativo ma dovrebbe essere positivo.
- **Reale negativo - True negative (TN)**: una situazione in cui sia l'istanza di predizione che l'istanza attuale di una classe sono negative.

Appartenenti a questa famiglia di definizioni di seguito vengono proposte alcune delle piú comuni [25, 26]:

- **Paritá Statistica – Statistical Parity**: Questa definizione richiede che la proporzione di risultati positivi (ad esempio, essere assunti per un lavoro, ricevere un prestito) sia la stessa per gruppi diversi (ad esempio, generi o razze diversi).
- **Paritá di Opportunitá – Equal Opportunity**: Questa definizione richiede che la previsione del tasso di reali positivi di un risultato sia lo stesso per gruppi demografici diversi, pur tenendo conto delle differenze di falsi positivi.

- **Uguaglianza del Trattamento – Treatment Equality**: Questa definizione richiede che i risultati previsti siano gli stessi per gruppi diversi, indipendentemente dal trattamento (ad esempio, ricevere un intervento o meno).

Il problema principale risiede nel fatto che le definizioni statistiche di Fairness producono risultati equi tra gruppi demografici protetti e non protetti ma dal punto di vista dell'individuo sono palesemente scorrette [27]. Le definizioni individuali di Fairness invece utilizzano coppie specifiche di individui piuttosto che effettuare un calcolo statistico su gruppi demografici diversi. Si basano sul principio che "individui simili dovrebbero essere trattati in modo simile", dove la somiglianza è definita rispetto ad una metrica particolare che deve essere determinata caso per caso. L'algoritmo, presi in input due individui con le stesse caratteristiche dovrebbe produrre predizioni e risultati simili, anche se appartengono a gruppi demografici differenti [25]. Le principali strategie di trattamento di bias algoritmici e di mitigazione delle problematiche etiche si suddividono in 3 tipologie [28]:

- **Pre-processing**, il dataset di allenamento viene modificato prima di effettuare il training del modello di ML. Questa operazione comporta la rimozione di attributi sensibili o il bilanciamento del dataset per assicurarsi della rappresentazione omogenea tra tutti i gruppi.

- **In-processing**, vengono applicati vincoli di fairness durante il processo di addestramento.

- **Post-processing**, i risultati dell'algoritmo vengono manipolati per soddisfarre i requisiti di fairness.

CAPITOLO 4

Metodologie di ricerca

L'obbiettivo di questo studio empirico consiste nell'individuazione di elementi in comune tra i concetti di privacy e fairness nell'ambito del machine learning. Essendo entrambi i concetti importanti vincoli qualitativi nella progettazione di un modulo di machine learning, viene posta particolare attenzione sull'influenza reciproca che privacy e fairness possono avere. Per raggiungere lo scopo di questo studio viene effettuata una Systematic Literature Review (SLR), un metodo di ricerca utilizzato per identificare, valutare e sintetizzare in modo sistematico la conoscenza attualmente presente relativa ad un argomento di interesse [29].

4.1 Quesiti di ricerca

Il primo passo per effettuare una SLR consiste nell'individuazione delle research question, una o piú domande specifiche che aiutano nell'identificazione dei concetti chiave da analizzare. Come spiegato precedentemente, lo scopo di questo studio é finalizzato alla ricerca di una relazione tra i concetti di privacy e fairness nel contesto del machine learning, dunque:

◎ **Obbiettivo:** Identificare le principali relazioni e implicazioni comuni tra i concetti di privacy e fairness nel contesto del machine learning.

Questo porta alla formulazione delle seguenti Research Questions (RQ):

RQ$_1$. *Esistono relazioni di dipendeza tra fairness e privacy nello sviluppo di soluzioni di machine learning?*

Questa prima domanda di ricerca mira a verificare l'esistenza di relazioni dirette tra i concetti di fairness e privacy, quali definizioni, proprietá o scelte specifiche di sviluppo che influenzino reciprocamente entrambe le proprietá non funzionali.

RQ$_2$. *In quali applicativi machine learning specific, le implicazioni e le dipendenze tra fairness e privacy sono particolarmente rilevanti?*

Rispondendo a questa domanda di ricerca, si vuole analizzare in quali ambiti applicativi, le relazioni tra fairness e privacy nello sviluppo ML, risultano essere particolarmente rilevanti, riportandone esempi concreti.

RQ$_3$. *Esistono, ad oggi, strumenti automatici atti a misurare, trattare in maniera congiunta le implicazioni dirette tra privacy e fairness nello sviluppo ML?*

Attraverso questa domanda viene posto come obbiettivo principale l'analisi del funzionamento dei tool attualmente sviluppati, qualora esistessero, per misurare l'impatto di soluzioni per la tutela della privacy sulle implementazioni fair-oriented e viceversa.

4.2 Query di ricerca

Nella seconda fase della revisione sistematica della letteratura vengono determinati i termini di ricerca chiave che possono aiutare a recuperare tutte le informazioni necessarie per condurre la ricerca. Vengono osservati i seguenti passaggi per la composizione della query di ricerca:

1. Per ogni domanda di ricerca sono state derivate le parole chiave piú rilevanti.
2. Per ogni parola chiave sono stati identificati i sinonimi principali.
3. É stato utilizzato l'operatore Booleano OR ($\vee$) per incorporare i sinonimi.
4. É stato utilizzato l'operatore Booleano AND ($\wedge$) per collegare le parole chiave tra di loro.

La query di ricerca risultante é la seguente:

("Privacy" $\vee$ **"Private")** $\wedge$ **("Fairness"** $\vee$ **"Fair")** $\wedge$ **("Machine Learning"** $\vee$ **"ML")**

4.3 Sorgenti di ricerca

Per poter selezionare i principali documenti scientifici sull'argomento di interesse é stata sottomessa la query nel motore di ricerca **Google Scholar**, che indicizza le risorse letterarie attualmente pubblicate sugli archivi di ricerca digitali. Per ottenere una panoramica piú ampia sugli articoli accademici sono state effettuate ulteriori ricerche sui seguenti database :

- **ArXiv** (`https://arxiv.org/`)
- **IEEEXplore** (`https://ieeexplore.ieee.org/`)
- **ACM Digital Library** (`https://dl.acm.org/`)

4.4 Criteri di selezione delle risorse

Una Systematic Literature Review prevede un processo di scrematura su tutto l'insieme di materiale accademico generato dalla sottomissione della query di ricerca negli archivi sopracitati. I criteri di esclusione e inclusione determinano i requisiti di selezione che ogni paper scientifico deve avere affinché sia appropriato per la ricerca da condurre.
Nelle seguenti tabelle vengono elencati i criteri di esclusione ed inclusione che sono stati adottati.

Criteri di esclusione
Data di pubblicazione precedente al 2015
Nome dell'autore non specificato
Articoli non redatti in inglese
Documenti duplicati
Articoli il cui testo non era accessibile
Studi al di fuori dell'ambito di ricerca
Fonte non specificata
Scarsa comprensibilitá
Materiale non accademico
Titolo del documento non contenente le parole chiave o sinonimi di "Fairness" e "Privacy"

Tabella 4.1: Criteri di esclusione

Per evitare una ridondanza dei record nella tabella relativa ai criteri di inclusione sono stati omessi i vincoli opposti ai criteri di esclusione. Ad esempio é implicito che se gli articoli non redatti in lingua inglese non sono presi in considerazione per la ricerca, i documenti scritti in inglese soddisfano i requisiti di selezione.

Criteri di inclusione
Studi che esaminano in dettaglio una relazione diretta tra privacy e fairness
Documenti che analizzano o sfruttano tool atti ad identificare relazioni tra privacy e fairness
Articoli che esplicitano ambiti di applicazione in cui esiste una relazione tra privacy e fairness

Tabella 4.2: Criteri di inclusione

4.5 Estrazione dei dati

Dopo aver identificato le risorse che nei risultati di ricerca dalla query utilizzata rispettavano i criteri di inclusione ed esclusione, al fine di avere una panoramica organizzata del lavoro da svolgere, le informazioni relative ai documenti sono state raccolte in una tabella. Nello specifico é stata definita una tabella denominata come "Data Extraction" all'interno della quale sono state inserite le informazioni piú rilevanti per rispondere alle domande di ricerca. I dati riportati nella tabella sono i seguenti :

- ID documento
- Informazioni bibliografiche
- Tipologia di studio
- RQ1
- RQ2
- RQ3

Per semplificare il processo di estrazione é stato associato ad ogni articolo un ID e nella colonna relativa alle informazioni é stato inserito un riferimento alle note bibliografiche, in modo da avere una panoramica completa su tutti i dati della risorsa. Per ogni colonna relativa alla research question é stato inserito un check (✓) per determinare l'idoneitá di un documento a rispondere a tale domanda. In questo modo é stato possibile avere una prospettiva chiara delle risorse a disposizione e identificare eventuali obbettivi comuni che ognuna di esse si poneva di affrontare.

4.6 Sintesi e combinazione dei dati

L'obiettivo della sintesi dei dati è quello di aggregare le prove provenienti dagli studi per rispondere alle domande di ricerca. I risultati e le conclusioni di un solo studio possono avere poco peso nel processo di evidenza o confutazione di una tesi, ma l'aggregazione di piú risorse puó aiutare ad ottenere una panoramica chiara su

un argomento di interesse. I dati sono stati analizzati e sintetizzati prendendo in considerazione solo la parte degli articoli che risponde alle domande di ricerca. In particolar modo per ogni documento si é posta maggiore attenzione su:

- Articoli nei quali é stata analizzata una relazione specifica tra definizioni di fairness e privacy.
- Articoli nei quali é fornita una panoramica dei contesti reali in cui vengono applicati moduli di machine learning dove é possibile analizzare il riscontro dell'influenza reciproca tra fairness e privacy.
- Articoli dove l'analisi della relazione in questione é stata supportata dall'implementazione di un tool che dimostrasse in modo chiaro l'influenza di soluzioni per la tutela della privacy sul concetto di fairness e viceversa.

I dati infine sono combinati e analizzati per poter rispondere alle research questions.

CAPITOLO 5

Analisi dei risultati

5.1 Esecuzione del processo di ricerca

Una volta definite le metodologie applicate per eseguire una Systematic Literature Review, si é proceduto alla sua esecuzione. In particolare il lavoro é stato diviso nei seguenti passaggi:

1. Sottomissione della query di ricerca nei database:
    - **Google Scholar**: 109.000 risultati.
    - **ArXiv**: 236 risultati.
    - **IEEEXplore**: 122 risultati.
    - **ACM Digital Library**: 9100 risultati.
2. Ogni elemento della ricerca é stato sottoposto all'applicazione dei criteri di esclusione. Durante questa fase ad ogni documento sono stati applicati i filtri dei criteri di esclusione, sono stati presi in considerazione il titolo, l'abstract e le parole chiave di ciascun articolo. Grazie all'applicazione dei vincoli si é potuta effettuare un'ottima scrematura sui risultati, riducendo considerevolmente le

risorse da prendere in considerazione. Dopo aver sottomesso la query di ricerca nei database ed aver applicato i filtri, i risultati sono stati i seguenti :

- **Google Scholar**: 144 risultati.
- **ArXiv**: 29 risultati.
- **IEEEXplore**: 4 risultati.
- **ACM Digital Library**: 9 risultati.

3. Ciascuno dei documenti rimanenti é stato sottoposto all'applicazione dei criteri di inclusione. A differenza del passaggio precedente, l'inclusione del documento é stata valutata prendendo in considerazione l'intero articolo, non solo il titolo, l'abstract e le parole chiave. Come risultato di questa procedura sono stati considerati 15 articoli.

4. Nella fase finale ogni documento é stato analizzato accuratamente, le sezioni da cui si potevano evincere risposte alle research questions designate sono state evidenziate per poterle studiare successivamente. In questa fase le informazioni relative ai documenti sono state estratte nella Extraction Data table. Le tabella risultante é la seguente:

Data Extraction					
ID	Informazioni Bibliografiche	Tipologia Studio	RQ1	RQ2	RQ3
P1	[30]	Sperimentale	✓		✓
P2	[31]	Sperimentale	✓		✓
P3	[28]	Sperimentale	✓		✓
P4	[32]	Sperimentale	✓		✓
P5	[33]	Sperimentale	✓		✓
P6	[34]	Sperimentale	✓	✓	✓
P7	[35]	Survey	✓	✓	
P8	[36]	Sperimentale	✓	✓	✓
P9	[37]	Sperimentale	✓	✓	✓
P10	[38]	Sperimentale	✓	✓	
P11	[39]	Sperimentale	✓	✓	✓
P12	[40]	Sperimentale	✓		✓
P13	[41]	Sperimentale	✓		✓
P14	[42]	Sperimentale	✓		
P15	[43]	Sperimentale	✓		✓

Tabella 5.1: Data Extraction Table

5.2 Analisi dei risultati ottenuti - RQ1

Per rispondere alla prima domanda di ricerca é stata effettuata un'indagine approfondita sui documenti per identificare quali metriche di privacy e definizioni di fairness fossero state studiate in modo intersezionale. Mediante quest'analisi é stato generato un grafico che riporta le metriche che sono state messe a confronto nei vari documenti. Per avere una panoramica concreta sui documenti analizzati, le definizioni di fairness sono state inserite nelle rispettive categorie di appartenenza: definizioni di gruppo o individuali. Viene fatta notare la presenza di una macro

area di definizioni di fairness denominata come "Fairness generica", all'interno della quale sono state inserite le definizioni generiche che sono state elaborate ad hoc per la problematica da risolvere nel relativo documento, o le definizioni che non sono state esplicitamente definite. Analogamente la macro area PPA (privacy preserving algorithms) racchiude i documenti in cui vengono utilizzati modelli di machine learning pre-addestrati con tecniche di tutela della privacy, dunque tali metodi di sicurezza non vengono implementati e formalmente definiti all'interno dell'articolo. All'interno delle risorse prese in esame vengono individuati i seguenti gruppi di confronto tra tecniche per garantire tutela della privacy e notazioni di fairness:

- **M1: DP e definizioni di fairness di gruppo**, la relazione di dipendenza tra il framework di privacy differenziale e le definizioni di fairness di gruppo é ampiamente discussa nei paper [30, 31, 28, 32, 33, 36, 40, 41, 43].

- **M2: PPA e definizioni di fairness individuali**, questa specifica relazione viene affrontata nel documento [37], gli algoritmi utilizzati sono TrajGAN e Mo-PAE, a cui vengono sottoposti test di performance da una prospettiva di fairness individuale. Viene elaborata una specifica metrica di somiglianza, relativa al problema dell'analisi delle traiettorie umane, per implementare la definizione di fairness individuale.

- **M3: PPA e definizioni di fairness generiche**, questa relazione viene discussa nei documenti [34, 39]. Nella ricerca condotta da Y. Resheff et al. [34] viene applicato un framework adversarial privacy su un algoritmo per i suggerimenti pubblicitari, la definizione di fairness non viene specificata. Lo studio condotto da S. Noiret et al. [39] definisce una definizione di fairness specifica per il problema da affrontare (bias nel riconoscimento facciale), chiamata Personal Detection Rate (PDR). Vengono utilizzati 4 classificatori a cui vengono applicate metodi di offuscamento delle immagini.

- **M4: DP e definizione di fairness generica**, il confronto intersezionale tra le due metriche viene indagato nei documenti [35, 42]. A. Xiang [35] non definisce esplicitamente la definizione di fairness adottata, mentre C. Tran et al. [42] definiscono una metrica di equità specifica che tiene conto della volontá di

equidistribuire le risorse tra varie entitá o organizzazioni. Un esempio concreto é dato dall'allocazione di fondi tra le scuole.

- **M5: Crittografia e definizione di fairness generica**, questa relazione viene affrontata nei documenti [35, 38], in entrambe le risorse la definizione di fairness non viene esplicitamente definita.
- **M6: Crittografia e definizioni di fairness gruppo**, questa relazione viene affrontata nell'articolo [43], il metodo di crittografia utilizzato si basa sulla secure multiparty computation (SMC).

Il grafico ottenuto é il seguente:

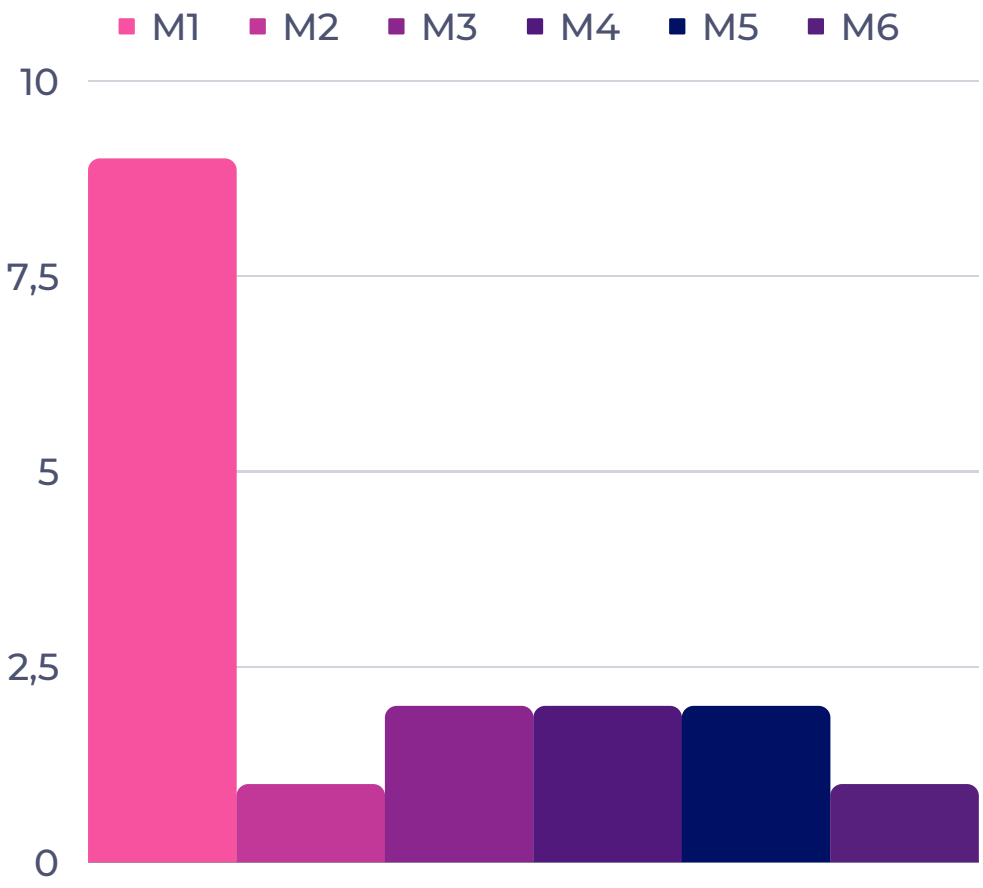

Figura 5.1: Grafico metriche privacy e fairness

Dal grafico si evince che la metrica di privacy che viene piú studiata in relazione al concetto di fairness é la privacy differenziale. Oltre la metá dei documenti [30, 31, 28, 32, 33, 36, 40, 41, 43] analizza in maniera congiunta le implicazioni dell'utilizzo

del framework di privacy differenziale con definizioni di fairness di gruppo.
Viene fatto notare che le definizioni di fairness meno studiate in relazione alle tecniche di tutela della privacy appartengono alla branca delle definizioni individuali di fairness. Come spiegato nella sezione 3.2.1 la difficoltá nell'utilizzo di definizioni di fairness individuali risiede nell'identificazione di un parametro di somiglianza che va stabilito caso per caso. Tra i documenti selezionati, soltanto lo studio condotto da Y. Zhan [37] effettua un'indagine su algoritmi per la tutela della privacy da una prospettiva di fairness individuale. La seguente tabella riporta in modo esaustivo le specifiche metriche di privacy e definizioni di fairness utilizzate nelle risorse studiate:

ID	Metriche Privacy	Definizioni Fairness
P1	Privacy Differenziale	Equalized Odds
P2	Privacy Differenziale	Equalized Odds
P3	Privacy Differenziale	Equal Opportunity
P4	Privacy Differenziale	Demographic Parity Equalized Odds Predictive Parity
P5	Privacy Differenziale	Demographic Parity Equal Opportunity
P6	PPA Mutual Information Privacy	Generica
P7	Privacy Differenziale Crittografia (FHE)	Generica
P8	Privacy Differenziale	Statistical Parity
P9	PPA (TrajGAN, Mo-PAE)	Demographic Parity Fairness individuale ad hoc
P10	Crittografia (ORE, PKE)	Generica
P11	PPA Gaussian Blur Pixelation	Generica
P12	Privacy Differenziale Locale	Disparate Impact Statistical Parity Equal Opportunity Overall Accuracy
P13	Privacy Differenziale	Demographic Parity
P14	Privacy Differenziale	Definizione ad hoc
P15	Privacy Differenziale Crittografia(SMC)	Disparate Impact Equal Opportunity Equalized Odds

Tabella 5.2: Tecniche di privacy e definizioni specifiche di fairness usate nei documenti.

Data la varietá di notazioni di fairness di gruppo utilizzate é stata ritenuta necessaria la creazione di un grafico a barre che prendesse in considerazione solo le definizioni analizzate in modo congiunto alla metrica di privacy differenziale, essendo lo scenario piú investigato nei documenti. Dalla tabella 5.2, prendendo in considerazione i documenti appartenenti al gruppo M1 [30, 31, 28, 32, 33, 36, 40, 41, 43], é stato possibile ricavare un nuovo grafico che mostra le definizioni di fairness di gruppo specifiche che sono state studiate in relazione al framework di privacy differenziale. Il grafico ottenuto é il seguente:

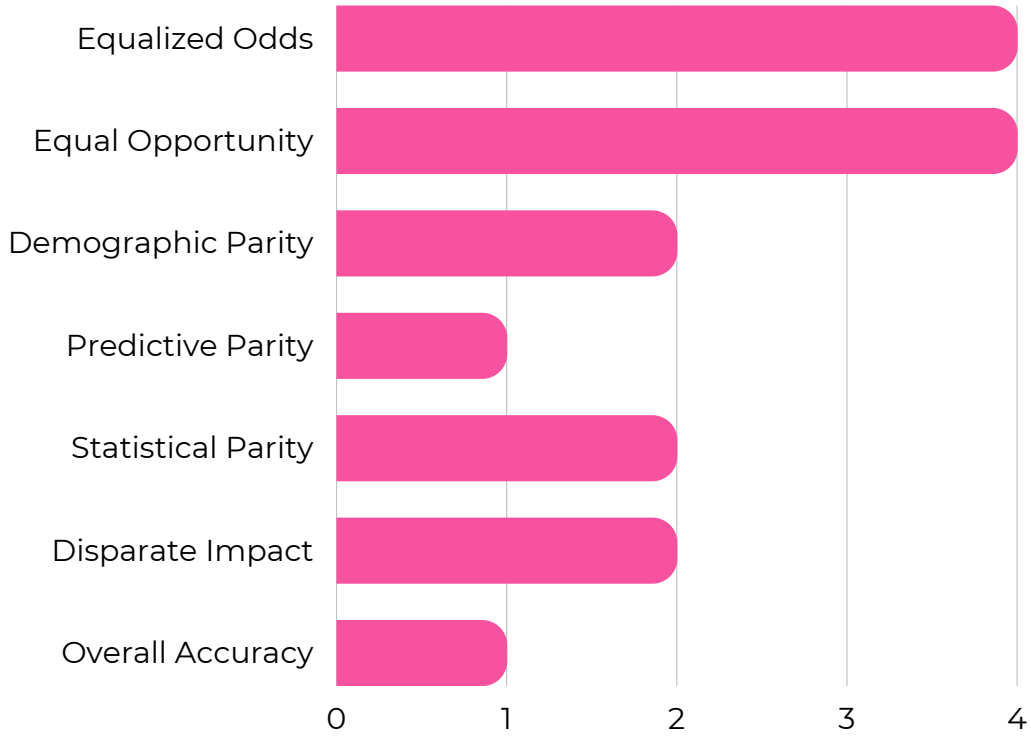

Figura 5.2: Grafico definizioni di fairness specifiche in relazione alla DP.

Si nota che le definizioni di fairness che sono state maggiormente analizzate attraverso un'ottica di privacy differenziale sono : Equalized Odds ed Equal Opportunity.

5.3 Analisi dei risultati ottenuti - RQ2

In questo paragrafo verrá fornita la risposta alla seconda domanda di ricerca. **In quali applicativi machine learning specific, le implicazioni e le dipendenze tra fairness e privacy sono particolarmente rilevanti?** Per dare una risposta a questa domanda sono stati esaminati i paper che analizzano la relazione tra i concetti di privacy e fairness all'interno di uno specifico contesto di applicazione di moduli machine learning. Dalle risorse selezionate é stato possibile individuare i seguenti settori in cui sono state riscontrate le implicazioni di una dipendenza tra privacy e fairness:

- **Sistemi di suggerimenti personalizzati - Recommender systems**
- **Computer Vision**
  - **Healthcare**
- **Smart Cities**

É importante sottolineare che in questo specifico caso di studio, il settore Healthcare é analizzato da una prospettiva della Computer Vision. L'unico documento relativo all'ambito dell'Healthcare ad aver soddisfatto i criteri di selezione é, infatti, la ricerca condotta da Arasteh et al. [36]effettua un'indagine sui trade-off tra privacy e fairness attraverso moduli di deep learning per il riconoscimento di immagini nel settore sanitario.

5.3.1 Recommender Systems

L'aumento della popolaritá nella fruizione di contenuti digitali ha fatto sí che i sistemi di suggerimenti personalizzati diventassero i maggiori influencer delle abitudini digitali. I recommender systems hanno rivoluzionato il modo in cui gli oggetti sono scelti in ogni contesto online, siano essi libri, film o prodotti da acquistare [34]. B. Mosher identifica due forme principali della pubblicitá digitale [38]:

- **Pubblicitá Mirata - Targeted Advertising**, in cui l'attivita dell'utente viene tracciata attraverso il web, e mediante tali informazioni é possibile offrire pubblicitá personalizzate.

- **Pubblicitá Contestualizzata - Contextual Advertising**, in cui i contenuti delle pubblicitá sono relative ai contenuti delle pagine web in cui vengono mostrate.

Le targeted advertising sono molto popolari sul web, anche se gli utenti hanno manifestato preoccupazioni circa la raccolta e diffusione di informazioni personali. La privacy, nell'ambito delle pubblicitá mirate, é una metrica difficile da bilanciare poiché l'utilizzo di protocolli di sicurezza avanzati puó inficiare sulla precisione delle predizioni dei modelli. Per poter effettuare un accurato matching tra i contenuti digitali e le preferenze dei clienti é necessaria una profilazione degli utenti. Nella maggior parte delle applicazioni reali le informazioni sugli utenti sono raccolte senza alcuna considerazione verso la privacy di questi ultimi. Dai profili degli utenti si possono ricavare informazioni riguardo gli interessi/hobby ma anche dati sensibili come il sesso, la religione o l'orientamento politico [38]. Nel contesto dei recommender systems, i documenti selezionati forniscono due prospettive diverse sulla relazione tra privacy e fairness. In particolare:

- Lo studio condotto da Resheff et al. [34] si concentra sull'analisi dei leak impliciti delle informazioni private nei recommender systems.
  Come dichiarato precedentemente da T. Mosher, i modelli di machine learning per le inserzioni pubblicitarie utilizzano sempre piú informazioni demografiche per poter generare predizioni piú accurate, creando al contempo minacce alla privacy. Gli autori dello studio, piuttosto che analizzare la problematica sfruttando framework di privacy differenziale, focalizzano l'attenzione sullo studio di tecniche di sicurezza atte a preservare dati che non sono stati esplicitamente forniti nei dataset di allenamento. Dando un esempio pratico, un attaccante che ottiene accesso alle informazioni relative al sesso di alcuni utenti puó utilizzare tecniche di supervised learning per ricavare il sesso di tutti gli altri utenti. La metrica di privacy utilizzata é comunemente nota come mutual information privacy (MI) e viene adoperata per quantificare i leak connessi alla privacy in un dataset.Nel documento viene proposto un metodo privacy-adversarial (resistente agli attacchi) per poter implementare recommender systems dai quali non si possono ricavare in maniera implicita dati sensibili. L'utilizzo di questa particolare metodologia per preservare la privacy ha delle implicazioni sul

concetto di fairness. Resheff et al. asseriscono che la soluzione al problema della discriminazione algoritmica risiede nella garanzia che un certo insieme di variabili (biased data o dati sensibili) non vada ad influenzare le predizioni di un modello. Gli autori sottolineano che escludere soltanto le informazioni sensibili dal modello non basta a garantire equitá. Viene posto l'esempio dell'esclusione di tutti i dati relativi al sesso per ottenere risultati equi nel rispetto di questo attributo. L'esclusione di queste informazioni non fornisce alcuna garanzia sul fatto che tali dati possano essere implicitamente ricavati da altri attributi, come ad esempio l'occupazione, poiché esiste una correlazione tra le variabili. A tal proposito applicando la metodologia di privacy-adversarial nel rispetto di un certo insieme di attributi protetti, ci si rende certi che nella rappresentazione del profilo di un utente tali dati non compaiano, garantendo predizioni imparziali e tutela della privacy. Nella ricerca viene dimostrato che, almeno nel contesto di recommender systems, il framework privacy-adversarial applicato al training di un modello funziona come teorizzato, anche utilizzando ulteriori classificatori non é possibile ricavare implicitamente informazioni relative al sesso o all'etá. Il metodo-adversarial puó essere utilizzato per minimizzare la presenza di qualsiasi dato sensibile noto durante il training. É interessante notare che irrealisticamente é possibile forzare l'esclusione di qualsiasi attributo, in tal caso peró non esisterebbero dati a sufficienza per poter effettuare suggerimenti personalizzati. Gli autori propongono l'utilizzo del framework per offuscare un piccolo insieme di informazioni sensibili, e utilizzare i restanti dati ricavati implicitamente per le pubblicitá personalizzate.

- A differenza dello studio condotto da Resheff et al. [34], T. Mosher contestualizza la dipendenza tra privacy e fairness all'interno del meccanismo che rende possibile l'assegnazione di una pubblicitá in uno spazio digitale: Real-Time Bidding (RTB). Dal documento si evince il seguente workflow relativo al RTB [38]:

  1. Un utente visita un sito web con spazio pubblicitario disponibile;
  2. Il sito web invia una richiesta ad un exchange pubblicitario, un'entitá responsabile degli accordi economici tra i proprietari del sito web(publisher),

che mettono in vendita uno spazio digitale, e inserzionisti. Un exchange pubblicitario é anche responsabile dell'associazione tra determinate proposte pubblicitarie e utenti. Nella richiesta sono contenute informazioni riguardanti l'utente e sullo spazio digitale disponibile.

3. Gli inserzionisti interessati allo spazio pubblicitario ricevono una richiesta d'asta.
4. L'inserzionista che propone la maggiore offerta, si aggiudica l'asta e lo spazio pubblicitario, la pubblicitá viene dunque mostrata all'utente.

Questo processo avviene in tempo reale, spesso in frazioni di secondo. Come spiegato precedentemente, per poter funzionare, i sistemi di advertising necessitano di una profilazione degli utenti, viene creata una rappresentazione astratta di un utente tramite le informazioni raccolte. Un possibile leak di queste informazioni puó rilevare strategie di marketing dei concorrenti, con conseguenti ripercussioni finanziarie. Una conseguenza analoga si puó riscontrare nel caso in cui l'offerta di un inserzionista in un'asta venisse divulgata, si puó ricavare ad esempio quanto un inserzionista offre su uno spazio pubblicitario, rivelando quindi strategie di marketing. Oltre alle sopracitate preoccupazioni riguardo la tutela della privacy degli inserzionisti ed utenti, ulteriore attenzione viene rivolta all'imparzialitá nell'assegnazione degli spazi pubblicitari. Per ogni asta lanciata esiste un rischio di sicurezza per il quale un exchange pubblicitario non valuti correttamente l'esito dell'asta. Nel contesto di sistemi pubblicitari, questo si traduce nel favoreggiamento di un inserzionista rispetto ad altri. La preoccupazione in questi sistemi é relativa alla fiducia intrinseca degli exchange pubblicitari. Gli inserzionisti possono basarsi solo sulla reputazione come metrica di fiducia, piuttosto che poter valutare autonomamente l'imparzialitá nella gestione delle aste. Come affermato anche nello studio di Resheff et al. [34] l'utilizzo di tecniche di privacy troppo stringenti, ad esempio l'irrealistica esclusione di ogni variabile relativa ad un utente tramite adversial-methods, puó compromettere l'abilitá di un modello di effettuare match efficienti tra inserzionisti e utenti. L'equitá tra inserzionisti é facile da ottenere senza considerare la privacy, le informazioni riguardo l'offerta di un'asta possono essere rese

pubbliche per convalidare i risultati dati dall'exchange. Tuttavia come spiegato precedentemente, la divulgazione di informazioni relative alle offerte di un'asta puó compromettere le strategie di marketing con forti impatti economici. In questo contesto la privacy é un fattore determinante per garantire sia l'equitá che il corretto funzionamento di un recommender system. L'autore della ricerca propone una soluzione per garantire sia privacy che fairness in un modello chiamato VIP-A.
VIP-A é un protocollo verificabile per la tutela della privacy per gli exchange pubblicitari (verifiable and privacy-preserving auction for ad exchanges). In altre parole é un metodo che garantisce sicurezza nella gestione delle aste permettendo agli inserzionisti di verificare se l'asta é condotta in modo imparziale rispetto alle informazioni fornite dall'exchange e dal proprietario del sito. Per poter consentire agli inserzionisti di verificare l'imparzialitá di un'asta pur non rivelando le offerte sottomesse all'exchange, vengono utilizzati protocolli di crittografia come metrica per garantire la privacy. In particolare VIP-A sfrutta la combinazione di due metodi di crittografia:

- **Order-revealing encryption (ORE)**, viene utilizzata per consentire agli inserzionisti di inviare le proprie offerte crittografate all'exchange pubblicitario, permettendo a quest'ultimo di effettuare confronti con altre offerte senza decifrarle.
- **Public-key encryption (PKE)**, viene utilizzata per tutelare la privacy dei prezzi delle offerte degli inserzionisti dai publisher.

T. Mosher afferma che VIP-A puó essere impiegato in recommender systems reali per garantire privacy ed imparzialitá, senza l'aggiunta di componenti esterne.

5.3.2 Computer Vision

Il rapido sviluppo dei sistemi di riconoscimento facciale é stato seguito da un crescente senso di ansia sulla sorveglianza di massa ad opera dell'intelligenza artificiale. Queste preoccupazioni hanno portato alla necessitá di un miglioramento nella tutela

della privacy e algoritmi piú equi [35]. I sistemi di sorveglianza non devono obbligatoriamente rappresentare una minaccia. In contesti come l'ambito sanitario sono indispensabili per monitorare in modo remoto la salute dei pazienti. In questi casi non é strettamente necessario conoscere l'identitá della persona nell'immagine. Algoritmi di preservazione della privacy delle immagini (privacy preserving algorithms-PPA) possono essere utilizzati, quindi, per tutelare la privacy del corpo [39]. Lo studio condotto da S. Noiret et al. [39] mira ad effettuare un'analisi sull'efficacia di PPA tramite una prospettiva di fairness. La ricerca esamina le discriminazioni basate sugli attributi protetti quali razza e sesso. La metrica di privacy adottata é l'offuscamento delle immagini, viene ottenuta attraverso due tecniche:

- **Sfocatura Gaussiana - Gaussian Blurring**;
- **Pixelizzazione - pixelation**.

I classificatori utilizzati per effettuare predizioni sono:

- **K-Nearest Neighbour (KNN)**;
- **Naive Bayes (NB)**;
- **Support Vector Classifier (SVC)**;
- **Multi-Layer Perceptron (MLP)**;

La definizione di equitá adottata é la fairness di gruppo. L'addestramento viene condotto su immagini non censurate e l'obbiettivo é la predizione degli individui da immagini censurate, viene valutata inoltre la capacitá di sfocare o pixelare in modo equo tutti i soggetti delle immagini. Quando viene utilizzata la tecnica di pixelizzazione, i gruppi che vengono riconosciuti con un tasso piú basso sono di sesso femminile e persone di etnia caucasica, in particolar modo donne bianche. Quando vengono considerati risultati incrociati tra i classificatori, il gruppo che viene riconosciuto con un tasso piú elevato sono donne non bianche. Quando viene utilizzata la tecnica della sfocatura, i gruppi che vengono riconosciuti con un tasso piú basso sono di sesso femminile e persone di etnia caucasica, in particolar modo donne bianche. Quando vengono analizzati i risultati intersezionali, i gruppi che ottengono i peggiori risultati

sono donne non bianche e uomini non bianchi. I risultati della ricerca condotta da S. Noiret et al. [39] evidenziano la presenza di discriminazione nell'offuscamento delle identitá sia utilizzando la pixelizzazione che la sfocatura. Gli autori asseriscono che non é stato possibile identificare l'origine del bias poiché il dataset era limitato. Nell'indagine condotta da A. Xiang [35] la relazione tra privacy e fairness viene definita come "tensione". Mentre i sistemi di riconoscimento facciale che tutelano la privacy mirano a limitare la raccolta di informazioni personali, la riduzione nelle discriminazioni algoritmiche necessita l'acquisizione di un voluminoso dataset variegato di immagini nitide per poter essere effettuata. Ció puó essere svantaggioso per i meccanismi di protezione della privacy, in quanto potrebbe comportare la raccolta di più informazioni personali di quelle che gli individui sono disposti a condividere. Nell'articolo vengono proposte alcune metodologie per bilanciare la tensione tra privacy e fairness:

- **Entitá esterne - Third-Part entites**, organizzazioni a cui viene attribuito un certo livello di fiducia per poter raccogliere informazioni personali in maniera etica e nel rispetto della privacy. Alcune di queste organizzazioni includono agenzie governative, istituzioni di ricerca e organizzazioni no-profit.

- **Privacy differenziale**, una sfida che risiede nell'utilizzo di questa tecnica é il bilanciamento dei benefici ottenuti attraverso l'offuscamento dei dati con gli svantaggi relativi all'equitá che si ottengono diminuendo la precisione sulle informazioni. Inoltre questo approccio non puó essere utilizzato su ogni tipologia di dati.

- **Homomorphic encryption**, é una tecnica che puó essere utilizzata per effettuare computazioni su dati criptati senza decifrarli. In questo modo si possono crittografare le immagini prima di essere inviate ad entitá di terze parti.

- **Synthetic Data**, nel contesto della computer vision, la generazione di dati sintetici puó essere utilizzata per la creazione di dataset di allenamento senza impiegare immagini reali. Puó essere implementata attraverso GAN (generative adversarial network) o modellizzazione 3D. La generazione di dati sintetici puó ridurre i livelli di discriminazione bilanciando piccoli dataset che contengono

poche informazioni sui sottogruppi demografici. La privacy viene garantita poiché non é necessario collezionare dati reali per l'addestramento dei modelli di machine learning. Un problema che puó sorgere utilizzando questa metodologia é l'introduzione di nuovi tipi di discriminazione dovuti all'impossibilitá di recepire totalmente la complessitá e le variabili di dati reali.

- **Federated Learning**, nell'ambito della computer vision puó essere utile in contesti dove gli individui sono riluttevoli nel condividere le informazioni personali con entitá di terze parti, ad esempio in ambienti finanziari.

L'autore del'indagine sostiene che ognuna delle metodologie sopracitate ha sia vantaggi che svantaggi nel bilanciamento della tensione tra privacy e fairness, perció si rende necessaria la combinazione di una o piú tecniche per ottenere risultati rilevanti.

5.3.3 Healthcare

Tra i documenti selezionati per effettuare la Systematic Literature Review, lo studio di Arasteh et al. offre un'analisi di un trade-off tra privacy e fairness in sistemi di machine learning applicati al contesto sanitario.
Il rapido sviluppo dell'intelligenza artificiale nel settore medico ha determinato un delicato compromesso. Da un lato i modelli devono offrire un'accurata precisione nella diagnostica trattando i pazienti in modo equo. Dall'altra, il personale sanitario é soggetto ad una responsabilitá etica e legale verso i dati dei pazienti utilizzati per l'allenamento dei modelli. In particolar modo, quando i modelli di diagnostica vengono condivisi con entitá di terze parti bisogna essere certi che la privacy dei pazienti non venga compromessa. Nel documento viene utilizzata la definizione di privacy differenziale come parametro di analisi e di confronto, infatti in letteratura risulta evidente l'inefficacia dei modelli di apprendimento distribuito per la tutela della privacy [36]. L'utilizzo della privacy differenziale come definizione di privacy pone le basi per l'analisi di un trade-off con il concetto di fairness. Arasteh et al. affermano intuitivamente che tramite l'offuscamento dei dati di allenamento tramite DP, un modello di machine learning apprende in maniera inversamente proporzionale i dati sui gruppi di pazienti sottorappresentati. In altre parole, la riduzione della precisione

sulle informazioni nei dataset di allenamento influisce in modo diretto sulle informazioni, giá a priori limitate, dei sottogruppi di pazienti. Il bilanciamento tra privacy e fairness in applicazioni sanitarie é un concetto molto delicato, le diagnosi errate non sono accettate cosí come le discriminazioni verso un certo gruppo di pazienti [36]. Gli autori della ricerca ritengono che per effettuare un'indagine approfondita sulla dipendenza diretta tra privacy e fairness in ambito medico, sia necessario testare il modello di machine learning in scenari qaunto piú affini alla realtá. Piuttosto che utilizzare dataset benchmark, come CIFAR-10 o ImageNet, viene impiegato un database clinico di radiografie pre indicizzate e un modello di machine learning privato. In contrasto con l'intuizione sopracitata, per cui l'applicazione della privacy differenziale crea proporzionalmente bias nelle predizioni, lo scopo principale della ricerca consiste nel dimostrare che é possibile ottenere diagnostiche accurate garantendo la privacy senza discriminazioni. Per la dimostrazione della tesi la stessa achitettura di classificazione (ResNet9) viene sottoposta sia ad un allenamento privato, tramite privacy differenziale, che non privato per poter analizzarne le differenze. La figura sottostante schematizza l'approccio utilizzato per la ricerca da Arasteh et al.

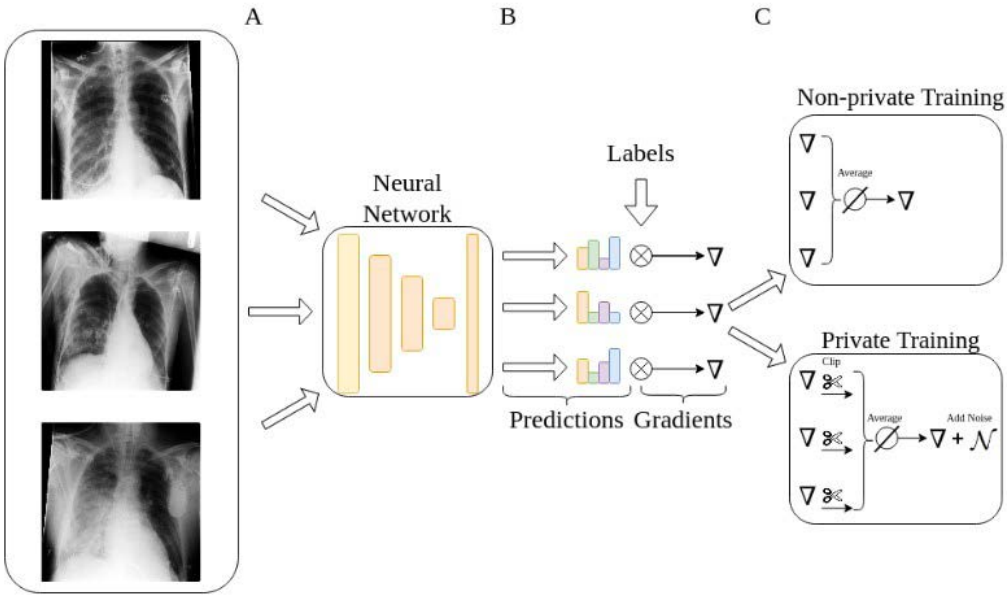

Figura 5.3: Processo di training privato e non privato di un modello di deep learning

Nello studio come metrica di fairness viene utilizzata una definizione di paritá statistica, mentre per analizzare la performance del modulo di machine learning

viene impiegata la metrica AUROC. I risultati della ricerca dimostrano che attraverso l'uso di sufficienti dati di esempio e design architetturali specifici, la presenza di una discriminazione verso alcuni sottogruppi non dipende dall'applicazione della privacy differenziale. I modelli studiati tendono ad effettuare diagnostiche imparziali pur garantendo la tutela della privacy, questo é sicuramente dovuto all'utilizzo di informazioni voluminose e di alta qualitá provenienti da contesti reali. É opportuno notare che nonostante l'applicazione o meno della DP non abbia implicazioni dirette sulla fairness, i modelli non sono comunque esenti dal commettere discriminazioni. Lo studio dimostra che i pazienti con etá superiore ai 70 anni sono soggetti a pregiudizi sia nel contesto di un modulo privato che non-privato, a differenza dei pazienti con etá inferiore ai 30 anni che sono meno soggetti a discriminazioni. Analogamente dall'analisi sulla fairness relativa al sesso del paziente si evince una precisione della diagnostica leggermente superiore nei pazienti di sesso femminile rispetto ai pazienti di sesso maschile. Per concludere la ricerca, Arasteh et al. asseriscono che, in questo specifico caso di studio, l'introduzione di meccanismi di privacy differenziale non amplificano pregiudizi relativi all'etá, al sesso o concomitanze patologiche dei pazienti [36].

5.3.4 Smart Cities

La comprensione della mobilitá umana basata sulle informazioni raccolte dai dispositivi mobili é diventata una componente fondamentale nella pianificazione di spazi urbani. Purtroppo,la raccolta di informazioni relative alla posizione degli spostamenti umani ha suscitato preoccupazioni inerenti alla privacy, soprattutto quando i dati contengono informazioni sensibili che possono rilevare informazioni sull'identitá, comportamenti, religioni delle persone [37]. Secondo Y. Zhan et al. [37] un punto di vista che é stato ampiamente trascurato, nell'applicazione di algoritmi orientati alla tutela della privacy, é se tali algoritmi funzionano correttamente per ogni individuo o se potrebbero portare a conseguenze imprevedibili nella protezione della privacy di un solo gruppo di persone. L'obbiettivo dello studio é la misurazione e la valutazione dell'imparzialitá degli algoritmi privacy-oriented applicati al tracciamento degli spostamenti umani. Nella ricerca vengono esaminati due modelli di

machine learning per la protezione della privacy:

- **TrajGAN**, é un modello di deep-learning end-to-end che genera dati sintetici che preservano caratteristiche spazio-temporali e attributi semantici sui dati di tracciamento della posizione.

- **Mo-PAE**, é un modello per la tutela della privacy che sfrutta un codificatore di informazioni. Mo-PAE addestra un codificatore a manipolare le rappresentazioni dei dati per trasmettere soltanto i dati utili alle predizioni. In questo modo si minimizzano le informazioni relative alle identitá degli utenti sfruttando tecniche adversarial-learning.

Entrambi i modelli sono basati sul paradigma PUT (privacy-utility trade-off), in cui l'obbiettivo é applicare protocolli di sicurezza senza inficiare sulle performance di un modello. In letteratura, questo tipo di approccio é risultato essere piú efficace rispetto all'utilizzo di framework come la privacy differenziale [37]. Per l'analisi del problema vengono applicate le definizioni di fairness di gruppo ed individuali. Come menzionato nella sezione 3.2.1, le definizioni individuali di fairness si basano sul principio che "Individui simili devono essere trattati in modo simile" dove la somiglianza é una metrica da stabilire caso per caso. In particolar modo si rende necessaria un adattamento della metrica di fairness individuale per poter essere studiata nel contesto dei sistemi di geolocalizzazione. Y. Zhan et al. [37] identificano due nozioni di somiglianza:

- SIM_t, basata sulla somiglianza strutturale delle traiettorie.

- SIM_o, basata sulla somiglianza tra i risultati dei modelli PUT.

Le definizioni di fairness di gruppo non sono strettamente dipendenti dal concetto di somiglianza e gli autori dello studio utilizzano una metrica standard (paritá demografica) cercando di minimizzare le differenze con l'applicazione a meccanismi privacy-oriented per l'analisi di dati spazio-temporali. I dataset utilizzati nella ricerca sono i seguenti:

- **MDC**, contenente informazioni registrate tra il 2009 e il 2011 da 184 di volontari nell'area di Losanna/Ginevra. MDC include informazioni individuali sensibili, quali etá, sesso e status occupazionale.

- **Geolife**, contenente 17.621 traiettorie raccolte da Microsoft Research Asia da 182 utenti tra il 2007 e 2011. Questo dataset non contiene attributi demografici, di conseguenza non é stato possibile misurare l'equitá da una prospettiva della fairness di gruppo.

Dopo aver applicato le metriche di fairness ai modelli di machine learning privati, si evince che la privacy individuale non é stata garantita in entrambi i modelli. Il modello TrajGan ottiene un tasso di discriminazione minore rispetto al modello Mo-PAE applicando una metrica di fairness individuale. Nell'analisi su entrambi i modelli PUT, non sono state riportate violazioni nei confronti della fairness di gruppo nel rispetto degli attributi quali sesso, etá e occupazione. Gli autori della ricerca sottolineano che i risultati ottenuti rispetto alla fairness di gruppo sono altamente influenzati dai dataset utilizzati. In particolare nel caso dell'MDC le informazioni sulle traiettorie raccolte sono relative al contesto socio-economico e alla libertá culturale associate allo stile di vita svizzero. Lo studio inoltre presenta limitazioni dovute alle scarse informazioni demografiche contenute nei dataset. I sistemi di geolocalizzazione privacy-oriented sono progettati in modo tale da limitare la diffusione verso gli attributi demografici protetti, i quali sono essenziali per condurre un'indagine sull'equitá. Questa prospettiva spiega anche il motivo per il quale i modelli PUT riescono a garantire fairness di gruppo ma non fairness individuale. Y, Zhan et al. concludono la ricerca sostenendo che il concetto di fairness é in stretta relazione con la privacy, ma la quantificazione di questa dipendenza é ancora poco chiara [37].

5.4 Analisi dei risultati ottenuti - RQ3

Per rispondere alla terza domanda di ricerca sono state individuate le tecniche adottate in ogni documento per misurare l'impatto di metodi per la tutela della privacy sulle nozioni di fairness. Sono state generate le seguenti tabelle in cui vengono riportate in maniera sintetica le informazioni relative all'algoritmo di machine learning utilizzato, i dataset e le conclusioni tratte circa le implicazioni di protocolli di privacy sulle definizioni di fairness.

ID	Modello / Algoritmo di ML	Dataset	Conclusioni
P1	Private- FairNR una versione che applica privacy differenziale all'agoritmo Fair-NR. La privacy differenziale viene garantita attraverso il meccanismo di Laplace per aggiungere "noise" alle subroutine del modello.	Non specificato	Il modello con un'alta probabilitá soddisfa in modo approssimativo i criteri di Equalized Odds fairness tramite meccanismi di DP. Non é possibile garantire equitá accurata tramite DP.
P2	Neural Network	Dataset generato con dati sintetici	I vincoli di fairness implicano iniquitá nella tutela della privacy dei sottogruppi. L'imparzialitá si puó ottenere al costo della privacy e l'utilizzo di vincoli di fairness puó portare a significativi leak di informazioni sensibili.
	Decision Tree con complessitá variabile	• COMPAS • LAW • BANK	
P3	Supervised learning model generico. La privacy differenziale viene garantita attraverso un meccanismo esponenziale.	• FICO • Dataset generato con dati sintetici	É possibile ottenere un classificatore imparziale utilizzanto un framework di privacy differenziale ma a costo dell'accuratezza nelle predizioni
P4	Algoritmo di deep learning addestrato tramite DP-SGD, una variante del tradizionale stochastic gradient descent algorithm a cui é stata applicata privacy differenziale.	• ACS • LSAC • ADULT • COMPAS	Applicare privacy differenziale a modelli di deep learning non necessariamente amplifica l'iniquitá. In alcuni casi puó ridurre le disparitá a seconda della nozione di fairness presa in esame.
P5	Algoritmo di deep learning addestrato tramite DP-SGD, una variante del tradizionale stochastic gradient descent algorithm a cui é stata applicata privacy differenziale.	CelebA artificialmente sbilanciato	L'utilizzo di protocolli stringenti sulla privacy inficia sull'accuratezza delle predizioni rendendo il modulo piú imparziale. Piú il comportamento dell'algoritmo diventa casuale piú si garantisce fairness.
P6	Privacy-Adversarial recommendation system, attraverso la metrica MI (Mutual information privacy) viene misurata la quantitá di leak di informazioni sensibili	MovieLens 1M	Il modello garantisce tutela della privacy in modo imparziale nel rispetto di determinati attributi sensibili.
P8	Algoritmo di deep learning (ResNet9) addestrato tramite privacy differenziale	UKA-CXR	L'applicazione di privacy differenziale non inficia sulla fairness.

Figura 5.4: Strumenti adoperati nei documenti selezionati

ID	Modello / Algoritmo di ML	Dataset	Conclusioni
P9	PPA basati su GAN: • Mo-PAE • Traj-GAN	• MDC • GEOLIFE	La fairness individuale non viene garantita dagli algoritmi considerati. Le metriche di fairness di gruppo non subiscono violazioni.
P11	Classificatori che sfruttano tecniche di pixelizzazione e sfocatura: • K-Nearest Neighbour (KNN) • Naive Bayes (NB) • Support Vector Classfier (SVC) • Multi-Layer Perceptron (MLP)	PUBFIG	Vengono riportate discriminazioni nei confronti di attributi quali sesso e razza. L'iniquitá non dipende dall'utilizzo di classificatori che sfruttano tecniche di offuscamento delle immagini
P12	LGBM addestrato tramite privacy differenziale locale	• ADULT • ACS • LSAC	L'utilizzo del framework di privacy differenziale locale non inficia significativamente sulle performance e non genera iniquitá
P13	Algoritmi di logistic regression (PFLR e PFLR*) a cui viene applicata privacy differenziale	• ADULT • DUTCH	Negli algoritmi proposti vengono garantiti sia i requisiti di privacy che fairness preservando la precisione delle predizioni.
P15	PrivFairFL, un framework di federated learning che combina privacy differenziale e SMC	• ADS • MovieLens-1K	La soluzione proposta risulta efficace per garantire fairness di gruppo pur preservando la privacy.

Figura 5.5: Strumenti adoperati nei documenti selezionati

Si puó notare come le conclusioni degli studi siano molto eterogenee tra loro, anche nelle ricerche in cui viene adoperato lo stesso protocollo di privacy (DP) i risultati sono spesso contraddittori. Ad esempio lo studio condotto da R. Shokri et al. [31] viene concluso asserendo che l'equitá viene garantita rinunciando alla tutela della privacy, mentre dalla ricerca condotta da S. Arasteh et al. [36] si evince che l'utilizzo di privacy differenziale non ha implicazioni sul concetto di fairness. Il motivo di queste contraddizioni risiede nei differenti metodi di approccio al problema. R. Shokri et al. [31] conducono un'indagine da due prospettive, l'intersezione tra privacy e fairness viene studiata sia sfruttando una rete neurale allenata su dati sintetici

che un albero decisionale addestrato su dataset benchmark (COMPAS,LAW,BANK). Dall'altra parte, gli autori della ricerca [36] si pongono con uno sguardo critico sulle metodologie piú popolari presenti nello stato dell'arte per misurare le implicazioni congiunte di privacy e fairness. S. Arasteh et al. [36], infatti, sfruttano un'architettura ad hoc di deep learning e dati reali su radiologie etichettate a mano dal personale sanitario.

Nonostante i differenti approcci alla problematica un'ipotesi che trova riscontro in piú documenti é che l'utilizzo di tecniche di privacy troppo restrittive possono inficiare a tal punto sulle performance del modello da renderlo inadatto a svolgere la task assegnata. Per questo motivo tutte le risorse elencate nelle figure 5.4 e 5.5 prendono in considerazione un ulteriore parametro, definito come utility o accuracy (precisione), che viene analizzato congiuntamente a privacy e fairness. T. Farrand et al. [33] e M. Khalili et al. [28] dimostrano empiricamente che esiste una relazione inversa tra la precisione di un modello e i leak di informazioni. Piú sono stringenti i vincoli di privacy applicati piú la precisione nell'elaborazione di risultati del modello diminuisce. Gli autori dell'articolo [33] asseriscono che se il valore di utility in un modello di machine learning é troppo basso, l'attivitá di classificazione diventa casuale a tal punto da non effettuare piú discriminazioni. R. Cummings et al. [30] dimostrano che non é possibile soddisfare contemporaneamente privacy, accuracy e fairness assoluta utilizzando un framework di privacy differenziale. Conseguentemente a dimostrazioni di questo tipo, la comunitá scientifica ha adottato nuovi paradigmi di sicurezza per poter bilanciare in modo efficace il trade-off tra privacy e utility. I modelli sviluppati nel rispetto di PUT (privacy-utility trade-off) sfruttano tecniche GAN o reinforcement learning piuttosto che adottare la privacy differenziale. Lo studio condotto da Y. Zhan et al. [37] infatti, utilizza a priori modelli basati sul paradigma PUT, chiamati Traj-GAN e Mo-PAE, ai quali vengono applicate metriche di fairness. Gli autori riescono a dimostrare che le definizioni di fairness di gruppo non vengono violate, garantendo tutela della privacy e l'efficienza dell'algoritmo.

Per dare una risposta netta alla domanda di ricerca in questione, dai documenti analizzati si evince che é possibile misurare le implicazioni dirette tra privacy e fairness. La misurazione coinvolge un'ulteriore metrica, ovvero la precisione nelle predizioni di un algoritmo di machine learning. Un modello soddisfa i requisiti di

equitá e privacy se é in grado di svolgere efficacemente la task assegnata. A causa delle limitazioni nella ricerca le evidenze ottenute sono relative al singolo caso di studio.

CAPITOLO 6

Conclusioni

Questa ricerca riporta una revisione sistematica della letteratura che mira a fornire una panoramica delle implicazioni che i framework di tutela della privacy hanno sul concetto di fairness. Dallo studio effettuato si nota che la privacy differenziale é ampiamente discussa in relazione alle definizioni di fairness di gruppo. Tuttavia l'eterogeneitá dei risultati ottenuti, non permette ad oggi di stabilire in modo universale se la privacy differenziale generi ulteriori iniquitá o se sia in grado di attenuarle. Un'ulteriore problema risiede nel bilanciamento del parametro utility, l'applicazione di tecniche di privacy troppo rigide puó compromettere le performance del modello a tal punto da renderlo inadatto allo svolgimento di una task. Lo studio congiunto tra i concetti di privacy e fairness nel machine learning é ancora una tematica poco affrontata in letteratura. I risultati delle ricerche presenti nella revisione sistematica sono spesso contraddittori.Le limitazioni al conseguimento della ricerca sono infatti molteplici, ad esempio la scarsa disponibilitá di dataset che contengono informazioni reali.A gravare sulla difficoltá di risoluzione di questo problema c'é l'ancora aperto dibattito su quale definizione di fairness utilizzare. Nello stato dell'arte infatti sono presenti oltre 20 definizioni di fairness e ancora non é stato possibile stabilire una notazione oggettiva ed universale. Questa problematica irrisolta non permette ai ricercatori di analizzare la tematica in tutto il suo spettro, bensí solo una piccola parte.

Uno sviluppo futuro potrebbe mirare ad analizzare le implicazioni di metriche di fairness individuali su modelli implementati tramite paradigmi PUT (privacy-utility trade-off). Questi ultimi infatti hanno dato risultati promettenti in quanto giá a priori garantiscono la utility del modello, mentre le definizioni individuali di fairness rappresentano ancora una difficoltá importante nella ricerca [37].

Ringraziamenti

Desidero esprimere la mia profonda gratitudine al Professore Fabio Palomba, relatore della mia tesi di laurea, e al Dott. Carmine Ferrara e Dott. Giammaria Giordano che con estrema disponibilitá, prontezza e dedizione mi hanno guidato in questo percorso finale del mio titolo di studio. Senza il vostro aiuto non sarei riuscito a raggiungere questo traguardo.

Bibliografia

[1] E. Alpaydin, *Introduction to Machine Learning, fourth edition*, ser. Adaptive Computation and Machine Learning series. MIT Press, 2020. [Online]. Available: https://books.google.nl/books?id=tZnSDwAAQBAJ (Citato alle pagine 3 e 4)

[2] R. Kohavi and F. Provost, "Glossary of terms," *Machine Learning*, vol. 2, pp. 271–274, 01 1998. (Citato a pagina 4)

[3] Wikipedia, "Computer — Wikipedia, the free encyclopedia," http://it.wikipedia.org/w/index.php?title=Computer&oldid=132829644, 2023, [Online; accessed 20-April-2023]. (Citato a pagina 4)

[4] "Science:the goof button," 1961. [Online]. Available: https://content.time.com/time/subscriber/article/0,33009,872691,00.html (Citato a pagina 4)

[5] J. M. Helm, A. M. Swiergosz, H. S. Haeberle, J. M. Karnuta, J. L. Schaffer, V. E. Krebs, A. I. Spitzer, and P. N. Ramkumar, "Machine learning and artificial intelligence: Definitions, applications, and future directions," *Current Reviews in Musculoskeletal Medicine*, vol. 13, pp. 69–76, 2020. (Citato a pagina 5)

[6] E. D. Cristofaro, "An overview of privacy in machine learning," *CoRR*, vol. abs/2005.08679, 2020. [Online]. Available: https://arxiv.org/abs/2005.08679 (Citato alle pagine 5, 7, 11 e 13)

[7] N. Papernot, P. D. McDaniel, A. Sinha, and M. P. Wellman, "Towards the science of security and privacy in machine learning," *CoRR*, vol. abs/1611.03814, 2016. [Online]. Available: http://arxiv.org/abs/1611.03814 (Citato alle pagine 5, 11 e 12)

[8] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *CoRR*, vol. abs/2011.11819, 2020. [Online]. Available: https://arxiv.org/abs/2011.11819 (Citato alle pagine 6, 7, 9, 10, 11, 12 e 13)

[9] M. T. Islam, A. Fariha, and A. Meliou, "Through the data management lens: Experimental analysis and evaluation of fair classification," *CoRR*, vol. abs/2101.07361, 2021. [Online]. Available: https://arxiv.org/abs/2101.07361 (Citato a pagina 8)

[10] K. M. Habibullah and J. Horkoff, "Non-functional requirements for machine learning: Understanding current use and challenges in industry," *CoRR*, vol. abs/2109.00872, 2021. [Online]. Available: https://arxiv.org/abs/2109.00872 (Citato a pagina 8)

[11] B. Liu, M. Ding, T. Zhu, Y. Xiang, and W. Zhou, "Adversaries or allies? privacy and deep learning in big data era," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 19, p. e5102, 2019, e5102 cpe.5102. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.5102 (Citato alle pagine 9 e 14)

[12] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers," *Int. J. Secur. Networks*, vol. 10, pp. 137–150, 2013. (Citato a pagina 10)

[13] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17. New York, NY, USA:

Association for Computing Machinery, 2017, p. 1175–1191. [Online]. Available: https://doi.org/10.1145/3133956.3133982 (Citato a pagina 10)

[14] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 3–18. (Citato a pagina 10)

[15] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan, "iprivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5. [Online]. Available: https://par.nsf.gov/biblio/10026310 (Citato a pagina 10)

[16] R. Shokri, M. Stronati, and V. Shmatikov, "Membership inference attacks against machine learning models," *CoRR*, vol. abs/1610.05820, 2016. [Online]. Available: http://arxiv.org/abs/1610.05820 (Citato alle pagine 12 e 13)

[17] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *CoRR*, vol. abs/1908.09635, 2019. [Online]. Available: http://arxiv.org/abs/1908.09635 (Citato alle pagine 14 e 16)

[18] A. Lambrecht and C. Tucker, "Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads," *Management Science*, vol. 65, pp. 2966–2981, 2019. (Citato alle pagine 14 e 15)

[19] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, "Saving face: Investigating the ethical concerns of facial recognition auditing," *CoRR*, vol. abs/2001.00964, 2020. [Online]. Available: http://arxiv.org/abs/2001.00964 (Citato a pagina 15)

[20] B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive inequity in object detection," *CoRR*, vol. abs/1902.11097, 2019. [Online]. Available: http://arxiv.org/abs/1902.11097 (Citato a pagina 15)

[21] Z. Obermeyer, B. W. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, pp. 447 – 453, 2019. (Citato alle pagine 15 e 16)

[22] N. A. Saxena, "Perceptions of fairness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 537–538. [Online]. Available: https://doi.org/10.1145/3306618.3314314 (Citato a pagina 17)

[23] C. Natali, *Etica nicomachea*, ser. BUR.. Classici greci e latini. Laterza, 1999. [Online]. Available: https://books.google.it/books?id=KkBxPgAACAAJ (Citato a pagina 17)

[24] S. Verma and J. S. Rubin, "Fairness definitions explained," *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1–7, 2018. (Citato alle pagine 17 e 18)

[25] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *CoRR*, vol. abs/1810.08810, 2018. [Online]. Available: http://arxiv.org/abs/1810.08810 (Citato alle pagine 17, 18 e 19)

[26] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," 2017. (Citato a pagina 18)

[27] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel, "Fairness through awareness," *CoRR*, vol. abs/1104.3913, 2011. [Online]. Available: http://arxiv.org/abs/1104.3913 (Citato a pagina 19)

[28] M. M. Khalili, X. Zhang, M. Abroshan, and S. Sojoudi, "Improving fairness and privacy in selection problems," *CoRR*, vol. abs/2012.03812, 2020. [Online]. Available: https://arxiv.org/abs/2012.03812 (Citato alle pagine 19, 28, 29, 30, 33 e 48)

[29] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – a systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, 2009, special Section - Most Cited Articles in 2002 and Regular Research Papers. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584908001390 (Citato a pagina 20)

[30] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern, "On the compatibility of privacy and fairness," ser. UMAP'19 Adjunct. New York, NY, USA: Association for Computing Machinery, 2019, p. 309–315. [Online]. Available: https://doi.org/10.1145/3314183.3323847 (Citato alle pagine 28, 29, 30, 33 e 48)

[31] H. Chang and R. Shokri, "On the privacy risks of algorithmic fairness," 2021. (Citato alle pagine 28, 29, 30, 33 e 47)

[32] A. S. de Oliveira, C. Kaplan, K. Mallat, and T. Chakraborty, "An empirical analysis of fairness notions under differential privacy," 2023. (Citato alle pagine 28, 29, 30 e 33)

[33] T. Farrand, F. Mireshghallah, S. Singh, and A. Trask, "Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy," *CoRR*, vol. abs/2009.06389, 2020. [Online]. Available: https://arxiv.org/abs/2009.06389 (Citato alle pagine 28, 29, 30, 33 e 48)

[34] Y. S. Resheff, Y. Elazar, M. Shahar, and O. S. Shalom, "Privacy and fairness in recommender systems via adversarial training of user representations," *CoRR*, vol. abs/1807.03521, 2018. [Online]. Available: http://arxiv.org/abs/1807.03521 (Citato alle pagine 28, 29, 34, 35, 36 e 37)

[35] A. Xiang, "Being 'seen' vs. 'mis-seen': Tensions between privacy and fairness in computer vision," *SSRN Electronic Journal*, 2022. (Citato alle pagine 28, 29, 30, 39 e 40)

[36] S. T. Arasteh, A. Ziller, C. Kuhl, M. Makowski, S. Nebelung, R. Braren, D. Rueckert, D. Truhn, and G. Kaissis, "Private, fair and accurate: Training large-scale, privacy-preserving ai models in medical imaging," 2023. (Citato alle pagine 28, 29, 30, 33, 34, 41, 42, 43, 47 e 48)

[37] Y. Zhan, H. Haddadi, and A. Mashhadi, "Analysing fairness of privacy-utility mobility models," 2023. (Citato alle pagine 28, 29, 31, 43, 44, 45, 48 e 51)

[38] B. T. Mosher, "Privacy and fairness for online targeted advertising," Ph.D. dissertation, Queen's University (Canada), 2022. (Citato alle pagine 28, 30, 34, 35 e 36)

[39] S. Noiret, S. Ravi, M. Kampel, and F. Florez-Revuelta, "Fairly private: Investigating the fairness of visual privacy preservation algorithms," 2023. (Citato alle pagine 28, 29, 39 e 40)

[40] H. H. Arcolezi, K. Makhlouf, and C. Palamidessi, "(local) differential privacy has no disparate impact on fairness," 2023. (Citato alle pagine 28, 29, 30 e 33)

[41] D. Xu, S. Yuan, and X. Wu, "Achieving differential privacy and fairness in logistic regression," in *Companion proceedings of The 2019 world wide web conference*, 2019, pp. 594–599. (Citato alle pagine 28, 29, 30 e 33)

[42] C. Tran, F. Fioretto, P. Van Hentenryck, and Z. Yao, "Decision making with differential privacy under a fairness lens." in *IJCAI*, 2021, pp. 560–566. (Citato alle pagine 28 e 29)

[43] S. Pentyala, N. Neophytou, A. Nascimento, M. De Cock, and G. Farnadi, "Privfairfl: Privacy-preserving group fairness in federated learning," *arXiv preprint arXiv:2205.11584*, 2022. (Citato alle pagine 28, 29, 30 e 33)