UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA BIOMEDICA

Bias e fairness nelle applicazioni cliniche dell'intelligenza artificiale

Relatore
Enrico Longato

Laureando
Zaccaria Fasoli

ANNO ACCADEMICO 2023 – 2024

Data di laurea 19/11/2024

Abstract

Negli ultimi anni si sta osservando un sempre maggior utilizzo di intelligenza artificiale (AI, *artificial intelligence*) nell'ambito clinico e con ciò si sono sollevate anche questioni riguardanti problemi di equità e pregiudizi che tali applicazioni possono avere. Questo elaborato mira ad esplorare come i *bias* nell'AI possano influenzare negativamente l'equità nelle decisioni cliniche, rendendo impari il trattamento di alcuni gruppi sociali, fornendo un'analisi approfondita delle cause dei *bias* e delle loro implicazioni reali nel contesto clinico. Comprendere da cosa vengono causati i *bias* e il loro impatto sulla sanità e sull'AI, è fondamentale per poter elaborare strategie risolutive e poter garantire un trattamento equo a ogni individuo.
Per raggiungere questi obiettivi, è stata effettuata una revisione critica della letteratura esistente, analizzando casi di studio concreti. L'elaborato si articola in cinque capitoli. Il primo capitolo introduce le problematiche trattate, fornendo una panoramica degli obiettivi dell'analisi. Il secondo capitolo approfondisce le basi dell'intelligenza artificiale, delineando le branche del *machine learning* e del *deep learning*, e concludendo con informazion i sulla diffusione dell'AI e i suoi utilizzi in ambito clinico. Il terzo capitolo descrive cos'è un *bias* e le principali tipologie che possono presentarsi in un algoritmo, per poi fornire una definizione di *fairness* e descrivere alcuni criteri comunemente usati per valutare l'equità di un algoritmo. Il quarto capitolo esamina l'impatto dei *bias* attraverso dati e *case study* scelti per dimostrare come i *bias* possano influenzare studi clinici reali, mostrando quattro studi distinti che analizzano tipologie diverse di *bias*. Infine, il quinto capitolo sintetizza i temi trattati nei capitoli precedenti, evidenziando i punti di maggior interesse.

Ringraziamenti

Un sentito ringraziamento va alla mia famiglia, alla mia mamma, al mio papà e a mia sorella, per il loro supporto incondizionato e per avermi sempre sostenuto e incoraggiato lungo questo percorso. Grazie di cuore ai miei amici, che con la loro vicinanza e il loro incoraggiamento hanno contribuito a farmi mantenere la determinazione necessaria per arrivare fin qui. Infine, ringrazio il mio relatore per la disponibilità, la pazienza e la preziosa guida nella stesura di questo elaborato, il cui supporto è stato essenziale per la realizzazione di questo lavoro.

INDICE

1. INTRODUZIONE

Molti settori stanno subendo una rivoluzione grazie all'intelligenza artificiale (AI, *artificial intelligence*) e al *machine learning* (ML). In particolare, il campo medico ha visto la nascita di strumenti avanzati per la diagnosi e il trattamento delle malattie. Questi strumenti consentono l'elaborazione di grandi quantità di dati, grazie alla loro elevata capacità di calcolo, migliorando l'efficienza dei sistemi clinici. Un esempio significativo è l'oncologia, dove, negli ultimi trent'anni, il tasso di mortalità è diminuito di circa il 32%. Questo miglioramento è reso possibile grazie alla continua ricerca in campo oncologico e all'evoluzione tecnologica. E tra le tecnologie in uso, l'AI si distingue permettendo, per esempio, di identificare schemi precedentemente sconosciuti o prevedere l'evoluzione di malattie un tempo poco comprese [1].

Tuttavia, come accade con qualsiasi strumento, un uso scorretto dell'AI può portare a risultati altrettanto errati. In particolare, gli algoritmi utilizzati in ambito clinico possono essere soggetti a *bias*, ossia pregiudizi o inclinazioni che influenzano le decisioni, spesso basati su preconcetti errati o su una valutazione inadeguata delle variabili coinvolte. In ambito clinico e tecnologico, il *bias* può derivare da dati rappresentativi di gruppi specifici o da interpretazioni soggettive e può presentare problemi di equità, soprattutto nelle diagnosi, nelle cure e nelle prognosi. Questo può avere ripercussioni sia sul personale medico che sui pazienti, con conseguenze negative per la salute. L'interesse per questi argomenti è cresciuto notevolmente negli ultimi anni, con il 70% degli articoli riguardanti *bias* e fairness pubblicati tra il 2022 e il 2023 [2].

Per affrontare queste problematiche, il Parlamento Europeo ha introdotto *l'Artificial Intelligence Act*, un atto legislativo che stabilisce normative per lo sviluppo e l'uso dell'AI. Tra i principi fondamentali di questa legge troviamo l'affidabilità, la sicurezza e la *fairness* dei sistemi di AI, aspetti cruciali per garantire un suo utilizzo etico ed equo [3].

Per trattare queste problematiche l'elaborato è strutturato in tre capitoli. Il primo fornisce un'introduzione al concetto di AI, con riferimenti al ML e al *deep learning* (DL), che verranno trattati nella sezione finale del capitolo. Il secondo capitolo affronta i concetti di *fairness* e *bias*, analizzando le tecniche per valutare l'equità di un algoritmo e discutendo l'origine dei *bias*, andando a elencare le tipologie principali di *bias*. Il terzo

capitolo si concentra sull'impatto dei *bias*, presentando *case study* tratti dalla letteratura scientifica per illustrare come tali distorsioni influenzano il sistema sanitario. Infine, nella conclusione, vengono riassunti i temi trattati e vengono indicate alcune tecniche utili per mitigare o risolvere alcuni problemi causa dalla presenza di *bias*.

2. FONDAMENTI DI INTELLIGENZA ARTIFICIALE E APPLICAZIONI CLINICHE

In questo capitolo si vogliono introdurre le basi dell'AI e le principali tecniche di apprendimento automatico che ne supportano l'evoluzione, con particolare riferimento al ML e al DL. Verranno esplorate le caratteristiche e i meccanismi principali di ML e DL, evidenziando come questi modelli siano in grado di analizzare enormi quantità di dati e, ad esempio, supportare il processo decisionale medico. Infine, saranno presentati dati e statistiche sull'attuale diffusione dell'AI nel settore sanitario, offrendo una panoramica delle sue applicazioni più promettenti e delle potenzialità per migliorare la diagnosi, il trattamento e la gestione delle patologie.

2.1. Intelligenza artificiale

L'AI può essere definita come la capacità di un sistema di replicare capacità umane quali il ragionamento, l'apprendimento, la pianificazione e la creatività [3]. Questo ramo delle scienze informatiche ha come obiettivo la creazione di sistemi ed algoritmi in grado di simulare l'intelligenza umana e, ad oggi, è uno dei settori tecnologici più integrati nella quotidianità, grazie alle numerose applicazioni che ne derivano.
Ciò è possibile perché, diversamente dagli algoritmi tradizionali, i sistemi di AI possono migliorare le proprie prestazioni attraverso l'apprendimento, elaborando e analizzando grandi quantità di dati. Questo costante miglioramento delle funzionalità tramite l'autoapprendimento differenzia l'AI dagli approcci informatici tradizionali, in cui le istruzioni sono rigide e predefinite. Inoltre, uno dei fattori che ha favorito la crescita di queste tecnologie negli ultimi anni è l'enorme quantità di dati disponibile online e la maggiore accessibilità della potenza di calcolo.
Tra le diverse tecnologie legate all'AI, il ML è un sottoinsieme che consente ai sistemi di "imparare" da set di dati precedentemente analizzati, migliorando la propria accuratezza senza essere esplicitamente programmati per ogni singolo compito. A sua volta, un ulteriore sottoinsieme del ML è il DL, che sfrutta le reti neurali profonde per identificare pattern complessi, spesso utilizzate in applicazioni come il riconoscimento di immagini o linguaggio naturale [4].

Esistono quindi diverse normative che mirano a promuovere lo sviluppo di algoritmi affidabili e sicuri, impedendo a enti o compagnie di utilizzare questo strumento in modo errato. Possiamo prendere come esempio l'Artificial Intelligence Act pubblicato dal Parlamento europeo l'08/06/2023, che mira a regolamentare l'uso dell'AI stabilendo linee guida per lo sviluppo di diversi tipi di algoritmi. Tra i sistemi che si intende vietare, l'atto cita i sistemi di credito sociale simili a quelli utilizzati nella Repubblica Popolare Cinese e i sistemi di riconoscimento facciale impiegati per categorizzare e controllare i cittadini. Un altro esempio sono i sistemi di manipolazione comportamentale, che in alcuni gruppi vulnerabili possono incoraggiare comportamenti pericolosi. L'atto menziona, ad esempio, giocattoli che potrebbero influenzare i bambini a commettere atti violenti [3].

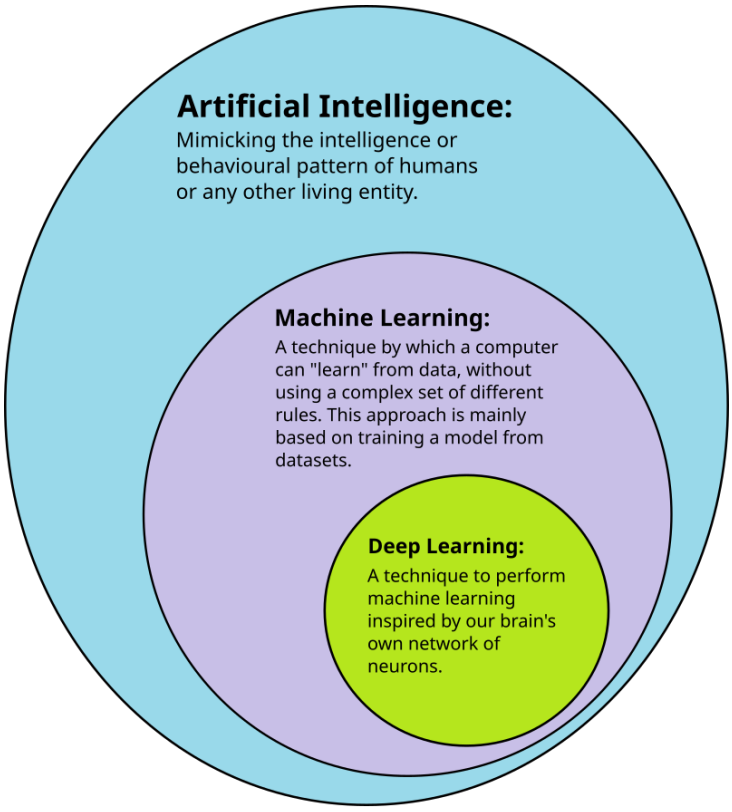

Figura 1: Il DL è un sottoinsieme del ML che è a sua volta un sottoinsieme dell'AI. Tratto da [5].

2.2. Machine learning

Il *ML* è un sottoinsieme dell'AI che consente ai computer di apprendere autonomamente senza essere esplicitamente programmati. Negli ultimi anni si è notato come gli sviluppatori comincino a preferire sistemi di ML addestrati mostrandogli esempi di dati di ingressi e uscite desiderati, permettendo ai modelli di migliorare le loro prestazioni grazie all'esperienza acquisita, piuttosto che programmare manualmente questi sistemi.
Il ML si distingue in diverse categorie principali, a seconda del tipo di dati e obiettivi di apprendimento. Il *supervised learning* è la forma più comune, in cui un modello viene addestrato su un set di dati etichettati, ovvero ogni esempio di ingresso ha associato un'uscita corretta. In contrasto, il *unsupervised learning* esplora pattern nascosti nei dati privi di etichette esplicite. Tecniche come la riduzione della dimensionalità e il clustering sono usate per trovare correlazioni implicite, sfruttando la struttura sottostante dei dati. Un esempio comune di applicazione è il clustering in ambito medico, dove i dati dei pazienti vengono raggruppati in base a somiglianze rilevanti senza informazioni pregresse sulle diagnosi.
Un'altra area significativa del ML è il *reinforcement learning*, categoria in cui gli agenti imparano interagendo con un ambiente e ricevendo feedback sotto forma di ricompense o penalità. Questo approccio si applica spesso nei sistemi di controllo autonomo, come nei robot o nei veicoli autonomi, dove le azioni di un agente influenzano l'ambiente e il feedback serve a migliorare progressivamente il comportamento [6][7].

2.3. Deep learning

Il DL è una branca del ML che utilizza reti neurali artificiali composte da vari layer per analizzare grandi quantità di dati e scoprire rappresentazioni utili senza la necessità di un intervento umano significativo. Ogni *layer* della rete neurale prende i dati in ingresso, li trasforma e invia il risultato al livello successivo, che ne ricava una rappresentazione più completa. Un'applicazione ben progettata del DL si distingue per la capacità di apprendere automaticamente queste rappresentazioni attraverso livelli

successivi di astrazione. Questo approccio si differenzia dai sistemi di ML tradizionali, che richiedono l'estrazione manuale di caratteristiche dai dati grezzi, come la trasformazione dei pixel di un'immagine in un formato compatibile con quello usato dall'algoritmo da addestrare.

Ad esempio, nel riconoscimento di immagini, i primi strati della rete individuano elementi semplici come bordi e contorni, mentre gli strati successivi combinano queste informazioni per identificare parti di oggetti e infine gli oggetti stessi. Un aspetto chiave del DL è l'uso dell'algoritmo di *backpropagation*, che permette di adattare dei pesi interni in base agli errori commessi durante il processo di apprendimento, migliorando così le sue prestazioni.

Il DL ha rivoluzionato settori come il riconoscimento vocale, il rilevamento di oggetti e la comprensione del linguaggio naturale, con applicazioni che spaziano dalla ricerca sul web ai sistemi di raccomandazione su piattaforme di e-commerce. Grazie alla disponibilità di grandi quantità di dati e alla crescente potenza di calcolo, è destinato a diventare ancora più importante nei prossimi anni [8][9].

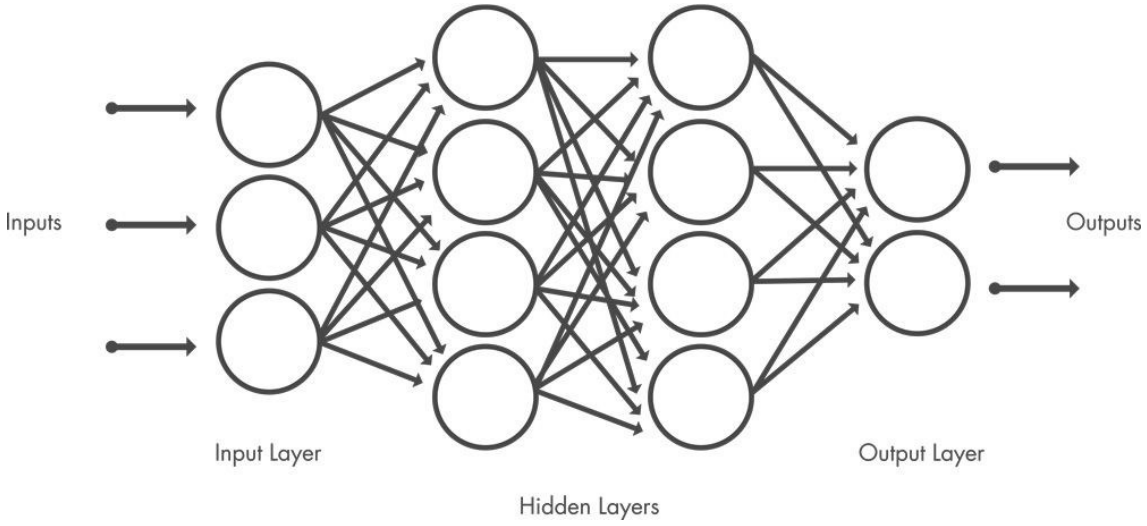

Figura 2: Stilizzazione di un CNN. L'immagine rappresenta la struttura semplificata di una CNN. I nodi rappresentano i neuroni, mentre gli archi, che in questa figura sono frecce, rappresentano i pesi. Il layer a sinistra, definito come input layer, riceve i dati in ingresso, e ogni nodo rappresenta una variabile o caratteristica dell'ingresso. I layer centrali, definiti come hidden layer, elaborano le informazioni ricevute dai nodi dell'input layer applicando i pesi; ogni nodo è connesso a quelli del livello successivo. Nell'ultimo layer, o output layer, ogni nodo rappresenta un valore di output [10].

Le reti neurali convoluzionali (CNN, *convolutional neural network*) rappresentano l'avanguardia nell'ambito del deep learning, in particolare per il riconoscimento delle immagini. Si basano su una struttura a strati simile alle reti neurali tradizionali, ma con una differenza chiave: i primi strati utilizzano operazioni di convoluzione per estrarre caratteristiche locali dai dati. Questa operazione consente di analizzare in dettaglio aree limitate di un'immagine, come bordi o angoli, che rappresentano informazioni rilevanti per la comprensione complessiva. In questo modo, i livelli successivi della rete possono combinare queste informazioni per riconoscere strutture più complesse e, infine, gli oggetti stessi. Una particolarità delle CNN è la capacità di mantenere invarianti alcune trasformazioni nell'immagine, come la rotazione o la scala, grazie a operazioni che riducono la complessità mantenendo intatte le informazioni essenziali. Questo rende le CNN particolarmente efficaci in applicazioni come il riconoscimento di volti e la classificazione di immagini.
Le Reti Neurali Ricorrenti (RNN, *Recurrent Neural Networks*) sono una classe di reti neurali artificiali progettate per elaborare sequenze di dati nel tempo, come il linguaggio naturale o segnali temporali. Grazie alla loro architettura ricorrente, le RNN possono "ricordare" informazioni precedenti all'interno della sequenza, rendendole particolarmente efficaci per compiti in cui il contesto è fondamentale [9].
I *transformer* sono modelli di DL che, per esempio, permettono di identificare relazioni e dipendenze tra parole in una frase, anche se lontane tra loro, migliorando significativamente la comprensione del contesto. Utilizzati principalmente per applicazioni di elaborazione del linguaggio naturale, i *transformer* sono alla base di molti sistemi avanzati, come i *chatbot* e gli strumenti di traduzione automatica [11].

2.4. Utilizzi ed impatto dell'AI sugli ambiti clinici

Grazie alla sua versatilità, l'AI trova una vasta gamma di utilizzi in ambito clinico, dove deve essere sottoposta a opportuni controlli e certificata da enti e normative che ne regolamentano l'uso. Uno degli enti più conosciuti e affidabili è la *Food and Drug Administration* (FDA), un'agenzia degli Stati Uniti d'America (USA, *United States of*

America), Paese che ha sviluppato circa il 30% dei sistemi AI approvati e attualmente in uso.

Fino al 25 giugno 2024, la FDA ha approvato 950 sistemi, di cui 221 solo nel 2023, che risulta essere l'anno con il maggior numero di sistemi approvati (Figura 3).

Approssimativamente il 90% dei sistemi approvati sono successivi al 2017; infatti, il 2018 sembra segnare una svolta e l'inizio di una crescita costante nel numero di sistemi approvati.

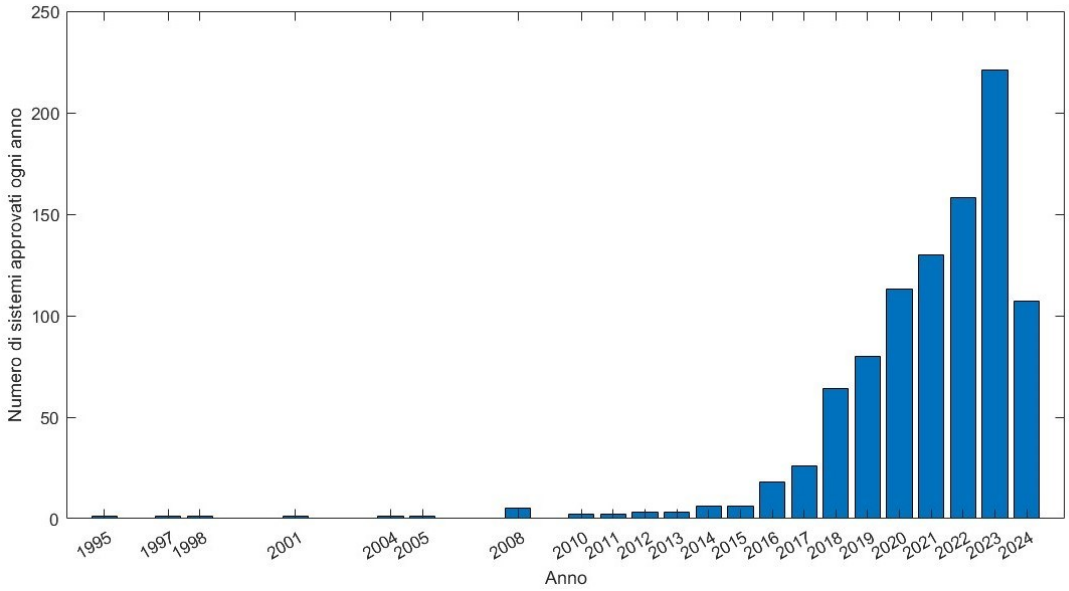

Figura 3: Sistemi AI-clinici approvati dall FDA dal 1995 al 2024. L'asse delle ascisse indica l'anno di approvazione mentre quello delle le ordinate indica il numero di sistemi approvati. Fonte [12].

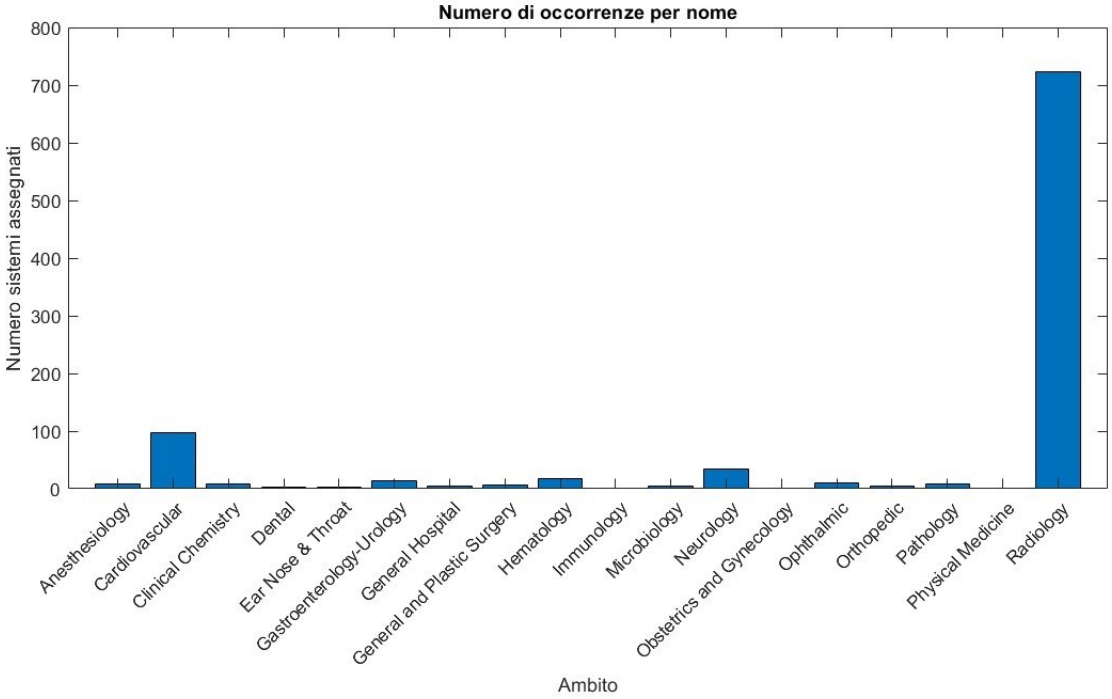

Figura 4: Diagramma rappresentante la quantità di sistemi approvati in ogni ambito. L'asse delle ascisse indica l'ambito di approvazione mentre quello delle le ordinate indica il numero di sistemi approvati. Fonti [12].

Si noti come il 76% dei sistemi approvati appartengono all'ambito della radiologia (Figura 4) con particolare attenzione ai sistemi di elaborazione immagini in particolare modo mirati al supporto oncologico [13]. Da ciò si può dedurre che l'AI sta rivoluzionando il campo della radiologia, migliorando l'efficienza e l'accuratezza delle diagnosi mediche. In particolare, l'utilizzo del DL ha permesso di automatizzare il riconoscimento di pattern complessi all'interno delle immagini radiologiche, compito tradizionalmente svolto dai radiologi. Questi algoritmi non solo offrono valutazioni quantitative che migliorano la precisione, ma consentono anche di ridurre errori legati alla soggettività umana [14]. Come anticipato un settore di grande interesse è l'oncologia, dove l'IA supporta l'identificazione precoce di masse tumorali, la caratterizzazione delle immagini tumorali e il monitoraggio delle risposte ai trattamenti. Per esempio, la radiomica basata su IA può prevedere la metastasi nei pazienti con carcinoma polmonare, offrendo un prezioso strumento per la stratificazione del rischio e la gestione clinica dei pazienti [15].

Inoltre l’AI sta trovando applicazioni significative anche in altri ambiti. Nella cardiologia viene impiegata per analizzare elettrocardiogrammi (ECG) ed ecocardiografia, rilevando anomalie come aritmie cardiache o ingrossamenti cardiaci. Nella gastroenterologia, l'AI è utilizzata per supportare i gastroenterologi nell’individuazione di polipi durante le colonscopie. L'AI può identificare polipi di dimensioni inferiori a 5 mm con un’accuratezza molto elevata, supportando diagnosi più rapide e accurate, anche in contesti di screening di routine. L'AI ha anche il potenziale di rivoluzionare la gestione delle malattie croniche, come il diabete e le patologie cardiache. Sistemi di monitoraggio continuo basati su algoritmi di ML possono analizzare i dati biometrici dei pazienti e prevedere complicazioni, consentendo un intervento precoce e migliorando la gestione complessiva della malattia [16].
I *chatbot* basati su AI stanno emergendo come strumenti innovativi nel supporto medico, offrendo assistenza continua ai pazienti e alleggerendo il carico dai professionisti sanitari. Questi sistemi, costruiti su modelli di linguaggio avanzati, come i modelli generativi di AI, possono fornire risposte alle domande dei pazienti, facilitare il triage e migliorare la comunicazione tra pazienti e medici, specialmente in contesti di telemedicina. Nonostante il potenziale, i *chatbot* affrontano alcune limitazioni, tra cui l’affidabilità e l’accuratezza delle informazioni fornite. Studi recenti hanno evidenziato che circa un terzo delle risposte dei *chatbot* non aderisce pienamente agli standard clinici, con il rischio di generare informazioni incomplete o poco chiare. Questo aspetto è particolarmente critico in campi come l’oncologia, dove la precisione delle informazioni influisce direttamente sulla fiducia dei pazienti e sulle decisioni terapeutiche [1].

3. *BIAS* E *FAIRNESS* NELL'AI

I *bias*, intesi come distorsioni sistematiche nei dati o negli algoritmi, possono condurre a decisioni discriminatorie, influenzando negativamente determinati gruppi di pazienti in base a caratteristiche come etnia, genere o status socioeconomico. Questo capitolo esplora le cause principali dei bias negli algoritmi di AI e analizza le diverse tipologie di distorsioni che possono sorgere nel processo di apprendimento automatico. Verranno inoltre introdotti i criteri fondamentali per valutare la *fairness* di un algoritmo, come l'indipendenza e il criterio di separazione e il criterio di sufficienza.

3.1. Cos'è un *Bias* e da cosa nasce

Il termine *bias* si riferisce generalmente a errori sistematici che si verificano durante un processo decisionale, portando a risultati ingiusti o non equi nei confronti di un individuo o di un gruppo specifico. Questi gruppi possono essere trattati in modo diverso a causa di fattori come etnia, età, genere, religione, cultura o condizione socioeconomica e tali discriminazioni non solo possono accentuare le disparità esistenti, ma rischiano anche di rafforzare stereotipi dannosi, alterando la percezione generale. Questa dinamica è particolarmente pericolosa perché rischia di esasperare le tensioni sociali, alimentando conflitti tra gruppi e perpetuando ingiustizie storiche.
Nel contesto dell'AI, i *bias* possono sorgere principalmente da due fattori distinti. In primo luogo, possono essere ereditati dai dati utilizzati per addestrare l'algoritmo. Se i set di dati di partenza sono già influenzati da pregiudizi o non rappresentano adeguatamente tutte le diverse popolazioni, l'AI imparerà e riprodurrà questi schemi di disuguaglianza. In secondo luogo, i *bias* possono derivare da difetti o imprecisioni nella progettazione dell'algoritmo stesso, che potrebbe riflettere ipotesi errate o criteri di decisione distorti [17].
Tuttavia, non tutti i *bias* sono necessariamente negativi. Alcuni *bias* sono essenziali, se non utili, per garantire che le diverse esigenze di individui di gruppi distinti siano prese in considerazione. Etnie diverse, sessi diversi e altre caratteristiche demografiche influenzano i bisogni di salute, e queste differenze devono essere evidenziate per assicurare che ogni persona riceva il trattamento più appropriato. Per esempio, le

differenze biologiche tra uomini e donne o le caratteristiche genetiche di diverse etnie possono influire sulla risposta ai farmaci o sui fattori di rischio per certe malattie. Ignorare queste distinzioni potrebbe portare a cure inefficaci o addirittura dannose. Questo rende la questione del *bias* molto più complessa, poiché è necessario un equilibrio tra il riconoscere e rispettare le differenze e il garantire che tali distinzioni non diventino discriminazioni.

Nel campo clinico, questa complessità è ancora più evidente. Oltre ai *bias* legati ai dati e agli algoritmi, emergono ulteriori forme di distorsione a causa della natura intricata delle interazioni umane e dei processi decisionali in ambito medico. L'eccessiva fiducia nei suggerimenti forniti da un sistema AI può portare a errori diagnostici, mentre i pregiudizi dei medici stessi, consci o inconsci, possono influenzare l'uso o l'interpretazione delle tecnologie AI [3]. Si deve inoltre considerare che gli utilizzatori dei sistemi di AI hanno aspettative e *bias* cognitivi quando generano e valutano i risultati di tali sistemi. Questi preconcetti influenzano la percezione e l'accettazione delle decisioni prese dall'AI, rendendo essenziale migliorare l'interazione tra l'AI e gli utilizzatori. Per rendere il sistema più familiare e trasparente agli utenti, è necessario introdurre elementi che favoriscano la comprensibilità delle decisioni. Come suggerito da studi nel campo della psicologia sociale e cognitiva, gli utenti tendono a richiedere spiegazioni che siano semplici, che rispondano alla domanda "Perché X invece di Y?". Questo approccio renderebbe i sistemi AI non solo più comprensibili, ma anche più affidabili, poiché le spiegazioni offerte rispecchierebbero le modalità con cui gli esseri umani generano e valutano le spiegazioni nella vita quotidiana [14]. Ciò considerando anche, la capacità di riconoscere le specifiche necessità legate a genere, etnia o altre variabili individuali che diventa cruciale per garantire un'assistenza sanitaria personalizzata e giusta per tutti.

3.2. Tipologie di *Bias* nell'AI

I *bias* si possono dividere in varie categorie in base a ciò che li origina, principalmente possiamo identificare i *data bias*, i *bias* da algoritmo, i *bias* d'interazione tra personale medico e AI e quelli d'interazione tra pazienti e AI (figura 5).

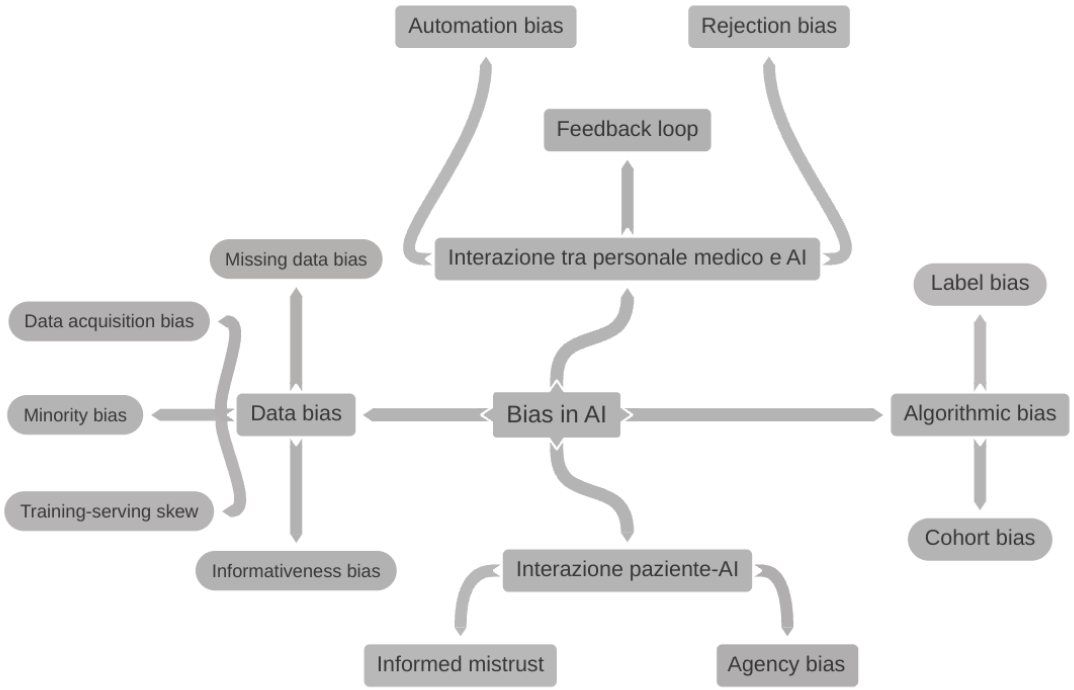

Figura 5: Bias nell'AI.
L'immagine illustra delle etichette introducendo varie tipologie di bias nell'AI per uso clinico [18].

Data bias è un termine che si riferisce all'incorretta organizzazione o raccolta dei dati usati per allenare l'AI e si dividono in alcune sottocategorie*, minority bias, missing data bias, informativeness bias* e Il *training-serving skew*.
I *minority bias* si riscontrano quando il numero di dati su una certa minoranza risulta essere insufficiente all'AI per imparare accurati schemi necessari per predire l'uscita del sistema. Ciò può portare a imprecisioni e risultati iniqui per i gruppi mal rappresentati.
I *missing data bias* vengono originati dalla mancanza di dati su un particolare gruppo rendendo impossibile per l'AI creare schemi predittivi. Possono essere visti come una versione estrema di *minority bias.*

L'*informativeness bias* si verifica quando le caratteristiche utilizzate per il rilevamento, pur essendo nominalmente le stesse, non riflettono equamente il valore per determinati gruppi protetti. Ad esempio, supponiamo che un algoritmo AI venga addestrato per prevedere il rischio di cancro al polmone. Se i dati di addestramento provengono principalmente da pazienti che si sono sottoposti a *screening* per sospetti pregressi, come nel caso dei fumatori, il modello potrebbe associare fortemente il rischio di cancro al polmone solo con i pazienti fumatori, ignorando la variabilità della popolazione generale. [19].
Il *training-serving skew* si riferisce alla discrepanza tra i dati utilizzati per l'addestramento dell'IA e quelli inseriti durante l'effettivo utilizzo. Questo può derivare da dati di addestramento non rappresentativi a causa di un *bias* di selezione o dall'utilizzo del modello su pazienti con una prevalenza nella popolazione diversa da quella dei dati di addestramento.
I *data acquisition bias* possono derivare dalle differenze nelle pratiche tra ospedali, come ad esempio nell'approccio alla raccolta dei dati o nella gestione dei protocolli clinici. Compromettendo la qualità e l'affidabilità dei dati provenienti da strutture diverse e limitandone la generalizzabilità [20].

Gli *algorithmic bias* nascono da un incorretto sviluppo e implementazione dell'AI e possono presentare *bias* a causa della loro logica, le principali sottocategorie sono i *label* e *cohort bias*.
I *label bias*, o *bias* di etichettatura, si viene a verificare quando l'addestramento dell'AI utilizza etichette incoerenti e che possono essere influenzate da disparità sanitarie. Questo può portare a decisioni distorte basate su informazioni imprecise o incoerenti negli algoritmi di AI. Ad esempio, è stato osservato un significativo *bias* razziale negli algoritmi disponibili in commercio utilizzati per prevedere i bisogni sanitari dei pazienti, oltre all'influenza di altri *bias*, uno dei fattori principali che ha contribuito a questo *algorithmic bias* è stato il suo design, che ha utilizzato il costo come indicatore dei bisogni sanitari, portando a una sottostima dei bisogni dei pazienti di certe etnie rispetto a quelli bianchi con condizioni simili.
I *cohort bias* sono causati quando l'AI è sviluppata usando modelli tradizionali e semplicistici che non prendendo in considerazione la granularità dei dati. Ad esempio, i

disturbi della salute mentale sono stati sottodiagnosticati o diagnosticati erroneamente nelle popolazioni lesbiche, gay, bisessuali, transgender, queer e altre (LGBTQ+). Una delle ragioni di ciò è che gli algoritmi spesso non tengono conto della complessità della popolazione LGBTQ+ e si basano solo su informazioni relative ai maschi e alle femmine biologici. L'AI addestrata su tali dati potrebbe continuare a trascurare o diagnosticare erroneamente i problemi di salute mentale in queste popolazioni, perpetuando potenzialmente le disparità esistenti nell'assistenza sanitaria mentale. esempio

I *bias* di interazione tra personale medico ed AI sono anch'essi vari, possiamo trovare gli *automation bias* ovvero la tendenza del personale medico a affidarsi ciecamente ai sistemi AI. Ad esempio i radiologi inesperti hanno una tendenza maggiore a seguire le indicazioni dell'AI senza controllare.
Un'altra condizione che può portare a *bias* è il *feedback loop*, che si ottiene quando il personale medico accetta le soluzioni proposte dall'AI anche se non corrette insegnando indirettamente all'algoritmo a perpetuare questi errori.
Il *rejection bias* può presentarsi quando il sistema supervisionato dal personale medico tende ad allertare in continuazione portando quindi il personale a perdere di fiducia nel sistema ignorando gli avvisi anche quando importanti.

I *bias* di interazione tra paziente ed AI sono i seguenti, il *privilege bias*, il quale si presenta quando una certa fetta di popolazione risulta impossibilitata ad accedere ad un certo sistema AI o a causa della scarsa distribuzione di macchinari ospedalieri portando all'aumento della disparità già presenti.
L'*informed mistrust* si riferisce allo scetticismo che alcuni gruppi possono avere verso l'AI o i sistemi clinici per motivi storici o pratiche non etiche passate. Ciò porta questi gruppi a non utilizzare i sistemi o direttamente le cure intenzionalmente.
L'*agency bias* è causata dalla mancanza di mezzi di comunicazioni o risorse di un gruppo che resta quindi ignorato durante lo sviluppo di un AI e quindi mal rappresentato da questo sistema [18].

3.3. Fairness nell'AI

La parola fairness è associata ai concetti di uguaglianza ed equità, valori fondamentali ripresi in molti ambiti, specialmente nelle costituzioni delle grandi democrazie, che considerano essenziale garantire l'equità tra i cittadini. Nella società, l'equità rappresenta un principio cardine per il funzionamento di un paese giusto e inclusivo.
Nell'ambito dell'AI, la fairness è un concetto complesso e oggetto di dibattito. Spesso, questo concetto viene associato all'assenza di *bias* o pregiudizi nelle decisioni degli algoritmi. Raggiungere, tuttavia, una vera equità in questi sistemi può risultare complicato, poiché le fonti di *bias* possono essere nascoste o strutturali, legate ai dati utilizzati per addestrare il modello o agli obiettivi che si cerca di raggiungere [17].
Esistono quindi diversi criteri comunemente usati per valutare la fairness di un algoritmo, e tra i principali troviamo: indipendenza, separazione e sufficienza.
Per esempio, poniamo il caso di una classificazione binaria dove $Y \in \{0,1\}$ rappresenta un'etichetta binaria usata per descrivere il modello e che, per compiti clinici, può riferirsi a stati oggettivi di salute, annotazioni soggettive o diagnosi, $R \in \{0,1\}$ denota il punteggio di classificazione $P(Y \mid X)$ assegnato dal nostro modello e si consideri A come attributo sensibile che definisce un dato sottogruppo protetto.

I. L'indipendenza, detta anche parità demografica, afferma che la frazione delle previsioni positive effettuate dal modello dovrebbe essere uguale tra i sottogruppi protetti, soddisfacendo il criterio di indipendenza $R \perp A$ tramite il seguente vincolo:

$$P(R = 1 \mid A = a) = P(R = 1 \mid A = b) \quad (1)$$

Per differenti sottogruppi a, b. L'indipendenza riflette l'idea che le decisioni dovrebbero essere prese indipendentemente dall'identità del sottogruppo. Tuttavia, si noti che la parità demografica vincola solo il tasso delle previsioni positive, e non considera il tasso con cui l'etichetta bersaglio effettiva può effettivamente verificarsi tra i sottogruppi. Immaginando un modello di AI che prevede la malattia renale. Se questo modello utilizza solo la parità demografica,

potrebbe diagnosticare la malattia renale ai pazienti afroamericani con la stessa probabilità con cui lo fa per altri gruppi, anche se questi pazienti possono avere un diverso rischio di malattia. Se, per esempio, consideriamo che la popolazione afroamericana ha un rischio effettivo più alto di malattia renale, ma il modello non lo considera e applica la parità demografica, potrebbe finire per sotto-diagnosticare i pazienti afroamericani. Oppure, se la malattia renale è meno comune tra i pazienti afroamericani, il modello potrebbe sovra-diagnosticare questo gruppo per mantenere la stessa percentuale di diagnosi positiva.

II. Il criterio di separazione o probabilità equalizzate afferma che i tassi di veri positivi e di falsi positivi dovrebbero essere uguali tra i sottogruppi protetti, soddisfacendo il criterio di separabilità $R \perp A \mid Y$ tramite i vincoli:

$$P\,(R = 1 \mid Y = 1, A = a\,) = P\,(R = 1 \mid Y = 1, A = b\,) \quad (2.1)$$
$$P\,(R = 1 \mid Y = 0, A = a\,) = P\,(R = 1 \mid Y = 0, A = b\,) \quad (2.2)$$

Rispetto all'indipendenza, la separabilità afferma che i punteggi degli algoritmi dovrebbero essere, ad opportune condizioni, indipendenti dall'attributo protetto, dato che l'etichetta bersaglio effettiva è nota. Pertanto, le probabilità equalizzate riconoscono che i sottogruppi potrebbero avere distribuzioni differenti della variabile target Y, e il loro obiettivo è ridurre gli errori in modo uniforme tra tutti i sottogruppi.

III. Il criterio di sufficienza o parità di qualità predittiva afferma che i valori predittivi positivi e negativi devono essere equi tra i sottogruppi, soddisfacendo il criterio di sufficienza $Y \perp A \mid R$ tramite il seguente vincolo:

$$P(Y = 1 \mid R = 1, A = a\,) = P(Y = 1 \mid R = 1, A = b\,) \quad (3.1)$$
$$P(Y = 1 \mid R = 0, A = a\,) = P(Y = 1 \mid R = 0, A = b\,) \quad (3.2)$$

Questo criterio assicura che i valori predittivi del modello siano uniformi tra i gruppi. Questo è particolarmente importante per garantire che la qualità delle previsioni sia equa per tutti i sottogruppi.

Quindi, questi criteri riflettono differenti approcci per garantire che i modelli di apprendimento automatico trattino equamente i gruppi protetti. Tuttavia, soddisfare tutti questi criteri contemporaneamente non è sempre possibile, e ciascuno ha i propri punti di forza e limiti a seconda del contesto applicativo.[21] Infatti emergono spesso conflitti tra diversi criteri di fairness. Questo è particolarmente evidente nel rapporto tra indipendenza e sufficienza, che tendono a escludersi reciprocamente. Infatti, se l'attributo sensibile A e la variabile target Y non sono indipendenti, non è possibile soddisfare simultaneamente entrambi i criteri. La sufficienza richiede che, dato il punteggio R, la variabile target Y sia indipendente dall'attributo sensibile A, mentre l'indipendenza implica che il punteggio non dipenda dall'attributo sensibile. Questa incompatibilità si estende anche al rapporto tra indipendenza e separazione, che richiede che la previsione R sia indipendente dall'attributo sensibile, dato il risultato effettivo Y. Pertanto, nel progettare sistemi equi, è importante considerare che l'adozione di uno di questi criteri potrebbe comportare la violazione di un altro, richiedendo un bilanciamento o compromessi specifici [22].

4. ANALISI DELL'IMPATTO DEI BIAS IN DIVERSI CONTESTI CLINICI E STUDIO DI CASI REALI

4.1. Impatto dei Bias nei Sistemi Clinici Basati su Intelligenza Artificiale

Nell'attuale sviluppo e implementazione di sistemi di diagnosi supportati da AI nella sanità, una problematica fondamentale è rappresentata dai *data bias*, ossia le distorsioni presenti nei dati utilizzati per l'addestramento degli algoritmi. La maggior parte dei sistemi di AI viene addestrata su banche dati che sovrarappresentano individui di discendenza europea o provenienti soprattutto da paesi occidentali. Questi dati non riflettono adeguatamente la diversità etnica globale, con il risultato che molte etnie, come afroamericani, ispanici e altre minoranze, non vengono adeguatamente rappresentate.

Tali sbilanciamenti etnici nei dati possono avere conseguenze significative. Un esempio è il cancro alla prostata, una malattia particolarmente aggressiva tra le popolazioni afroamericane e di discendenza ovest-africana. Le ricerche hanno dimostrato che la differente aggressività del tumore in questi gruppi può essere legata a variazioni genetiche specifiche, così come alle risposte biologiche divergenti a fattori ambientali e alimentari, inclusa la carenza di vitamina D, più comune tra persone con pelle scura. Questo esempio evidenzia come il genotipo e le condizioni ambientali possano influire diversamente sulla progressione di malattie gravi, ma l'uso di modelli predittivi basati su dati sbilanciati potrebbe sottovalutare tali variabili cruciali.

Gli algoritmi addestrati su popolazioni non rappresentative non solo rischiano di escludere le minoranze dalla possibilità di diagnosi accurate, ma possono anche introdurre disparità nei trattamenti. Ad esempio, i modelli AI progettati per diagnosticare malattie potrebbero utilizzare inconsapevolmente correlazioni spurie legate a fattori come l'etnia o lo status socioeconomico, portando a errori di valutazione clinica.

Le ricerche dimostrano che la rappresentazione inadeguata delle minoranze etniche nei *dataset* biomedici limita la capacità di scoprire variazioni genetiche significative, peggiorando le disuguaglianze già esistenti. Per esempio, la mutazione del Recettore del

fattore di crescita dell'epidermide (EGFR, *epidermal growth factor receptor*), rilevante nel trattamento del cancro ai polmoni, è stata scarsamente documentata nel "*The Cancer Genome Atlas* (TCGA)" per le popolazioni asiatiche, compromettendo l'efficacia delle terapie per i pazienti di origine asiatica.
L'inequità nei dati sanitari non si limita alla genetica, ma si estende a variabili come le condizioni socio-economiche e l'accesso alle cure. Gli algoritmi di ML sviluppati in contesti occidentali tendono a non tener conto delle determinanti sociali della salute, che includono fattori come il reddito, il livello di istruzione, l'accesso ai servizi sanitari e il contesto ambientale. Questi fattori giocano un ruolo centrale nel determinare la progressione e gli esiti di molte patologie, ma la loro influenza viene spesso trascurata nei modelli di AI, che possono quindi fallire nel fornire diagnosi o trattamenti adeguati ai pazienti provenienti da contesti meno privilegiati [23][24].

4.2. Analisi di casi reali

In questa sezione verranno analizzati casi reali che illustrano l'impatto concreto dei *bias* nei sistemi di AI. Attraverso questi esempi, verrà mostrato come errori sistematici nei dati possano influenzare le decisioni cliniche e contribuire a disuguaglianze nel trattamento sanitario. L'obiettivo è fornire una visione più approfondita delle implicazioni dei *bias* nei modelli AI, con particolare attenzione all'ambito medico.

4.2.1. Bias Raziale nell'Allocazione delle Cure: L'Impatto dell'Uso dei Costi Sanitari come Indicatore di Rischio

Nel caso di uno studio recente pubblicato [25], è stato analizzato un algoritmo utilizzato su larga scala nel sistema sanitario statunitense per la gestione dei pazienti a rischio. I pazienti sono stati etichettati in base a ciò che avevano autoproclamato di essere, dividendo poi le etnie tra due scaglioni, afroamericani e bianchi. Lo studio evidenzia quindi un *label bias* in cui l'algoritmo utilizza un'etichetta che non riflette accuratamente la condizione sanitaria del paziente.

Lo studio ha evidenziato un forte *bias* razziale: a parità di punteggio di rischio predetto dall'algoritmo, i pazienti afroamericani risultano significativamente più colpiti da varie malattie rispetto ai pazienti bianchi, infatti per esempio i pazienti afroamericani sono colpiti dal 26.3% in più di malattie croniche rispetto ai bianchi. Nonostante questo, i pazienti afroamericani ricevono meno cure. Questo fenomeno è attribuibile al fatto che l'algoritmo utilizza come etichetta principale il costo delle cure sanitarie, anziché basarsi direttamente sulle condizioni di salute del paziente. Poiché i pazienti afroamericani spesso ricevono meno assistenza medica a causa di barriere sistemiche, i costi sanitari risultano inferiori rispetto ai pazienti bianchi, nonostante una maggiore complessità clinica.
L'uso di etichette inappropriate come i costi sanitari anziché indicatori clinici oggettivi costituisce un esempio lampante di come i *bias* nei dati possano influenzare in modo distorto l'output di un algoritmo. In questo caso, il *bias* riflette una profonda disparità sistemica nel sistema sanitario americano, dove l'accesso diseguale alle cure viene riprodotto e amplificato dalle decisioni algoritmiche. I pazienti afroamericani, a parità di gravità della malattia, hanno quindi meno probabilità di essere identificati per ricevere cure aggiuntive, peggiorando ulteriormente le disuguaglianze già esistenti.
Questo tipo di *bias* non solo ha un impatto negativo immediato sui pazienti afroamericani, ma contribuisce a perpetuare le disuguaglianze razziali a lungo termine. Correggere tale distorsione migliorerebbe significativamente l'accesso alle cure per i pazienti afroamericani, aumentando dal 17.7% al 46.5% il numero di individui che riceverebbero assistenza aggiuntiva.

4.2.2. Bias di Rappresentazione nelle Immagini Cliniche: Implicazioni per la Diagnosi del Cancro della Pelle

Un altro studio esplora l'uso di reti neurali profonde per la classificazione del cancro della pelle, dimostrando che un modello di AI può raggiungere un livello di accuratezza comparabile a quello di dermatologi esperti [26]. Il sistema è stato addestrato su un vasto database di oltre 129.000 immagini cliniche di lesioni cutanee, rappresentando più di 2.000 malattie differenti. Tuttavia, nonostante i risultati promettenti, è emersa una

preoccupazione legata ai *bias* presenti nei dati di addestramento, in particolare, si possono evidenziare il *data bias* come *minority bias* e del *training-serving skew*.
Le immagini utilizzate per addestrare il modello provengono principalmente da popolazioni caucasiche, il che introduce un potenziale *bias*. Le differenze nella pigmentazione della pelle e nelle caratteristiche visive delle lesioni tra diverse etnie possono rendere il modello meno efficace nel diagnosticare correttamente il cancro della pelle in pazienti di origine non caucasica. Questo tipo di *bias* nei dati potrebbe portare a una sottorappresentazione delle varianti cliniche delle malattie nelle persone con pelle scura, compromettendo l'accuratezza diagnostica per queste popolazioni. Il rischio di questa incorrettezza è particolarmente preoccupante in un contesto clinico, dove la precisione e l'equità nelle diagnosi sono cruciali per la prevenzione e il trattamento delle malattie. Un modello addestrato prevalentemente su immagini di pazienti caucasici potrebbe non essere in grado di generalizzare correttamente per identificare il melanoma in persone con caratteristiche cutanee diverse, aumentando così il rischio di diagnosi errate o mancate.
Inoltre, le immagini utilizzate per addestrare il modello provengono da fonti molto standardizzate, come ospedali e cliniche altamente specializzati, il che potrebbe limitare la capacità del modello di funzionare in ambienti più variabili, come ambulatori o luoghi con risorse mediche limitate. Il *bias* presente nei dati di addestramento può quindi avere un impatto diretto sulla capacità del sistema di essere adottato su larga scala, in particolare in contesti dove la popolazione e le condizioni ambientali sono molto diverse rispetto a quelle rappresentate nel dataset di origine.

4.2.3. Bias di Generalizzazione nelle Radiografie: L'Influenza della Prevalenza Ospedaliera nella Diagnosi della Polmonite

Lo studio [27] ha esaminato le prestazioni di un modello addestrato con radiografie provenienti da tre diversi ospedali: il National Institutes of Health, il Mount Sinai Hospital e l'Indiana University. I risultati mostrano chiaramente che il modello, pur ottenendo buone performance quando testato su dati dello stesso ospedale in cui è stato addestrato, non riesce a generalizzare altrettanto bene su dati provenienti da ospedali esterni. Questo riflette un *bias* nei dati di addestramento che limita la capacità del

modello di fornire prestazioni coerenti su diverse popolazioni di pazienti, in particolare, questo può essere un altro esempio di *training-serving skew* e di *data acquisition bias.*
Il *bias* principale riscontrato in questo studio deriva dalla diversa prevalenza della polmonite nei tre ospedali. Ad esempio, la prevalenza della polmonite al Mount Sinai era molto più alta rispetto a quella degli altri due ospedali, il che ha permesso al modello di distinguere le immagini provenienti da ciascun ospedale con grande precisione, utilizzando informazioni falsate come il contesto ospedaliero piuttosto che concentrarsi esclusivamente sulle caratteristiche cliniche delle immagini. Ciò dimostra che il modello non stava imparando solo a rilevare la polmonite, ma anche a identificare inconsapevolmente il luogo di origine delle radiografie.
Questi *bias* possono avere conseguenze negative significative. Un modello che non riesce a generalizzare correttamente tra ospedali rischia di fornire diagnosi meno accurate quando utilizzato in ambienti clinici diversi da quelli in cui è stato addestrato. Ad esempio, pazienti provenienti da ospedali con una minore prevalenza di polmonite potrebbero ricevere diagnosi erronee, con conseguenti trattamenti inadeguati o ritardi nella cura. Questo dimostra che, nonostante i progressi della tecnologia di deep learning nella diagnosi medica, è fondamentale garantire che i modelli siano addestrati su dati diversificati e rappresentativi di una vasta gamma di contesti clinici.

4.2.4. Bias Etno-Clinico nel Carcinoma Polmonare: Impatti della Rappresentazione Inadeguata nei Dati sulla Diagnosi e Trattamento

Un altro esempio significativo dell'impatto che i *bias* nei dati clinici possono avere sui sistemi di AI è fornito dallo studio PIONEER, un'indagine epidemiologica condotta su pazienti asiatici affetti da carcinoma polmonare non a piccole cellule (NSCLC, *non–small-cell lung cancer*) [28]. Lo studio ha analizzato la frequenza delle mutazioni del EGFR, una mutazione cruciale che influenza la risposta ai trattamenti con inibitori delle tirosin-chinasi (EGFR-TKI, *epidermal growth factor receptor-tyrosine kinase inhibitors*) nei pazienti con adenocarcinoma polmonare in stadio avanzato.
I risultati dello studio hanno evidenziato che la frequenza di mutazione EGFR varia notevolmente tra i gruppi etnici e le caratteristiche cliniche. Ad esempio, la frequenza complessiva di mutazioni EGFR nei pazienti asiatici era significativamente più alta

rispetto a quella osservata nelle popolazioni caucasiche, con un valore del 51,4% tra i pazienti asiatici contro circa il 20% nelle popolazioni bianche. Inoltre, lo studio ha rilevato che le mutazioni EGFR erano più comuni nelle donne (61,1%) e nei non fumatori (60,7%), confermando differenze significative anche in base al sesso e allo status di fumatore.

Questo dimostra come l'assenza di una rappresentazione adeguata delle diverse etnie nei dataset clinici possa portare a modelli di AI che non sono in grado di generalizzare correttamente per gruppi di pazienti diversi. Nel caso specifico del carcinoma polmonare, l'uso di modelli sviluppati su dati raccolti prevalentemente da pazienti caucasici rischia di compromettere la precisione diagnostica e l'efficacia del trattamento nei pazienti di altre etnie, come gli asiatici, che presentano una maggiore incidenza di mutazioni EGFR.

Lo studia evidenzia un *minority bias* mostrando come questo tipo di *bias* non sia soltanto teorico, ma abbia implicazioni reali nella pratica clinica. Modelli di AI sviluppati senza una rappresentazione adeguata delle diverse popolazioni possono portare a trattamenti inadeguati o addirittura a diagnosi errate. Lo studio PIONEER sottolinea l'importanza di condurre test genetici, come quello per le mutazioni EGFR, su tutte le popolazioni, indipendentemente dal loro gruppo demografico o clinico di appartenenza. Ciò garantisce che il trattamento personalizzato, come l'uso di inibitori EGFR-TKI, possa essere offerto a tutti i pazienti che ne possono beneficiare, riducendo le disparità sanitarie e migliorando la *fairness* complessiva del sistema.

4.2.5. Commento ai casi di studio

Emerge chiaramente come i bias presenti nei dati clinici possano influenzare profondamente l'affidabilità e l'equità dei modelli di AI impiegati nel settore sanitario. In tutti i casi riscontriamo una problematica comune ossia la mancanza di rappresentatività dei dati di addestramento. Questa lacuna porta a modelli che non riescono a generalizzare adeguatamente su popolazioni diverse da quelle dominanti nei dataset originali, con conseguenze che spaziano dalle diagnosi imprecise alle disuguaglianze nei trattamenti.

In particolare, tutti gli studi evidenziano che il data bias si manifesta attraverso pratiche di raccolta dati limitate o non bilanciate, sia in termini di provenienza geografica, di etnia, sia di criteri socio-economici. Gli algoritmi, quando addestrati su dati raccolti in modo non inclusivo, rischiano di perpetuare e amplificare disparità preesistenti, un fenomeno osservato in tutti e quattro i casi.

Ciò che distingue i quattro studi, tuttavia, è l'origine specifica dei bias e le implicazioni cliniche che ne derivano. Ad esempio, nello studio sulle radiografie e nello studio PIONEER, il bias è legato alle caratteristiche della popolazione di riferimento e alle pratiche di raccolta dati nei diversi ospedali o gruppi demografici, configurando un *data acquisition bias*. Invece, nel caso della gestione del rischio dei pazienti, il problema principale risiede nel *label bias*, poiché si utilizzano etichette inappropriate come i costi sanitari, che riflettono disparità sistemiche anziché condizioni cliniche oggettive. Nella classificazione delle lesioni cutanee, il bias risulta invece dalla sovrarappresentazione di immagini di pazienti caucasici, un *minority bias* che mette in luce la difficoltà di addestrare modelli accurati e inclusivi per una popolazione eterogenea.

In sintesi, sebbene i bias emersi condividano un'origine comune nella raccolta o etichettatura non rappresentativa dei dati, ogni caso evidenzia diverse sfumature e impatti sul piano clinico. Questi esempi sottolineano la necessità di ampliare la rappresentatività dei dati clinici e di adottare criteri di addestramento più equi per promuovere l'affidabilità e la *fairness* degli algoritmi sanitari, riducendo le disuguaglianze nelle diagnosi e nei trattamenti per tutte le popolazioni.

5. CONCLUSIONE

I *bias* e la *fairness* sono problematiche che impattano in modo significativo sui sistemi informatici usati nella sanità, aggravando le disparità sociali già esistenti. Dall'analisi dei dati raccolti, emerge come siano principalmente le minoranze etniche a risentirne in ambito clinico, poiché la maggior parte dei dataset deriva da paesi con una maggioranza caucasica [24].

È importante sottolineare che, per eliminare i *bias* nei dati o rendere un sistema equo, non basta rimuovere le etichette che distinguono i gruppi; le differenze tra gli individui sono, infatti, ciò che permette un trattamento migliore e personalizzato. Quindi per poter mitigare l'impatto dei *bias* nell'AI, è fondamentale affidarsi a database più diversificati e rappresentativi delle minoranze. Questo consentirebbe di incorporare dati accuratamente selezionati, rendendo il sistema più flessibile e capace di gestire una maggiore varietà di casi clinici. Inoltre, è cruciale valutare e monitorare periodicamente gli algoritmi per garantire che rispettino le normative vigenti e che eventuali errori di programmazione non compromettano la fairness del sistema. Questa periodicità è necessaria anche per eliminare algoritmi obsoleti, considerato il rapido avanzamento tecnologico. Formare sia il personale clinico che i pazienti sull'uso degli strumenti basati su AI è indispensabile; è essenziale insegnare a fare affidamento su tali strumenti, ma senza accettare ciecamente ogni risultato prodotto dal sistema [18].

Inoltre, l'utilizzo dell'AI nei contesti clinici richiede una riflessione critica sulla fairness, considerata non solo come assenza di *bias*, ma come garanzia di un trattamento equo per tutti i pazienti. Come evidenziato negli esempi precedenti, le disparità demografiche insite nei dati utilizzati per addestrare i modelli di AI possono avere effetti profondi sulle decisioni sanitarie automatizzate, trasmettendo pregiudizi storici e creando un circolo di disuguaglianze. Il rischio di applicare decisioni predittive in ambito medico senza considerare queste implicazioni etiche può rendere invisibili alcune discriminazioni strutturali e perpetuarle involontariamente. Ogni criterio di *fairness*, dall'indipendenza alla sufficienza statistica, riflette intuizioni morali diverse e ha limitazioni nel rappresentare una fairness completa in ambito sanitario. Alla luce di questi fattori, è essenziale considerare strategie non solo per limitare l'impatto dei *bias*

esistenti ma anche per promuovere attivamente la *fairness* in tutti i livelli dell'applicazione clinica dell'AI [17].
Considerando le differenze tra gli individui, l'adozione di tecniche come quelle descritte e l'impiego di algoritmi equi e privi di *bias* potrebbe portare a un sistema sanitario migliore, in cui ogni persona venga trattata in modo equo, senza discriminazioni legate alle etichette, ma tenendo conto dell'unicità di ciascun caso.

6. BIBLIOGRAFIA

[1] L. Kolla e R. B. Parikh, «Uses and limitations of artificial intelligence for oncology», *Cancer*, vol. 130, fasc. 12, pp. 2101–2107, giu. 2024, doi: 10.1002/cncr.35307.

[2] F. Chen, L. Wang, J. Hong, J. Jiang, e L. Zhou, «Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models», *J. Am. Med. Inform. Assoc.*, vol. 31, fasc. 5, pp. 1172–1183, apr. 2024, doi: 10.1093/jamia/ocae060.

[3] «Artificial Intelligence Act» https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.

[4] M. Haenlein e A. Kaplan, «A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence», *Calif. Manage. Rev.*, vol. 61, fasc. 4, pp. 5–14, ago. 2019, doi: 10.1177/0008125619864925.

[5] Avimanyu786, "Immagine su Wikimedia Commons", CC BY-SA 4.0. [online]. Disponibile: https://creativecommons.org/licenses/by-sa/4.0. Data di accesso: novembre 9, 2024.

[6] F. Jiang *et al.*, «Artificial intelligence in healthcare: past, present and future», *Stroke Vasc. Neurol.*, vol. 2, fasc. 4, pp. 230–243, dic. 2017, doi: 10.1136/svn-2017-000101.

[7] M. I. Jordan e T. M. Mitchell, «Machine learning: Trends, perspectives, and prospects», *Science*, vol. 349, fasc. 6245, pp. 255–260, lug. 2015, doi: 10.1126/science.aaa8415.

[8] Y. LeCun, Y. Bengio, e G. Hinton, «Deep learning», *Nature*, vol. 521, fasc. 7553, pp. 436–444, mag. 2015, doi: 10.1038/nature14539.

[9] J. Schmidhuber, «Deep learning in neural networks: An overview», *Neural Netw.*, vol. 61, pp. 85–117, gen. 2015, doi: 10.1016/j.neunet.2014.09.003.

[10] MathWorks, "Convolutional Neural Network", MathWorks. [online]. Disponibile: https://it.mathworks.com/discovery/convolutional-neural-network.html. Data di accesso: novembre 9, 2024.

[11] M.-H. Guo *et al.*, «Attention mechanisms in computer vision: A survey», *Comput. Vis. Media*, vol. 8, fasc. 3, pp. 331–368, set. 2022, doi: 10.1007/s41095-022-0271-y.

[12] U.S. Food and Drug Administration, "Artificial Intelligence and Machine Learning (AI/ML) Enabled Medical Devices", FDA. [online]. Disponibile: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. Data di accesso: novembre 9, 2024.

[13] G. Joshi, A. Jain, S. R. Araveeti, S. Adhikari, H. Garg, e M. Bhandari, «FDA-Approved Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices: An Updated Landscape», *Electronics*, vol. 13, fasc. 3, p. 498, gen. 2024, doi: 10.3390/electronics13030498.

[14] A. Barredo Arrieta *et al.*, «Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI», *Inf. Fusion*, vol. 58, pp. 82–115, giu. 2020, doi: 10.1016/j.inffus.2019.12.012.

[15] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, e H. J. W. L. Aerts, «Artificial intelligence in radiology», *Nat. Rev. Cancer*, vol. 18, fasc. 8, pp. 500–510, ago. 2018, doi: 10.1038/s41568-018-0016-5.

[16] E. J. Topol, «High-performance medicine: the convergence of human and artificial intelligence», *Nat. Med.*, vol. 25, fasc. 1, pp. 44–56, gen. 2019, doi: 10.1038/s41591-018-0300-7.

[17] E. Ferrara, «Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies», *Sci*, vol. 6, fasc. 1, p. 3, dic. 2023, doi: 10.3390/sci6010003.

[18] D. Ueda *et al.*, «Fairness of artificial intelligence in healthcare: review and recommendations», *Jpn. J. Radiol.*, vol. 42, fasc. 1, pp. 3–15, gen. 2024, doi: 10.1007/s11604-023-01474-3.

[19] J. Kang *et al.*, «Artificial intelligence across oncology specialties: current applications and emerging tools», *BMJ Oncol.*, vol. 3, fasc. 1, p. e000134, gen. 2024, doi: 10.1136/bmjonc-2023-000134.

[20] K. Drukker *et al.*, «Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data

collection to model deployment», *J. Med. Imaging*, vol. 10, fasc. 06, apr. 2023, doi: 10.1117/1.JMI.10.6.061104.

[21] R. J. Chen *et al.*, «Algorithm Fairness in AI for Medicine and Healthcare», 24 marzo 2022, *arXiv*: arXiv:2110.00603. Consultato: 14 ottobre 2024. [Online]. Disponibile su: http://arxiv.org/abs/2110.00603

[22] S. Barocas, M. Hardt, e A. Narayanan, «Fairness and Machine Learning».

[23] C. Sudlow *et al.*, «UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age», *PLOS Med.*, vol. 12, fasc. 3, p. e1001779, mar. 2015, doi: 10.1371/journal.pmed.1001779.

[24] V. A. Zavala *et al.*, «Cancer health disparities in racial/ethnic minorities in the United States», *Br. J. Cancer*, vol. 124, fasc. 2, pp. 315–332, gen. 2021, doi: 10.1038/s41416-020-01038-6.

[25] «science.aax2342».

[26] A. Esteva *et al.*, «Dermatologist-level classification of skin cancer with deep neural networks», *Nature*, vol. 542, fasc. 7639, pp. 115–118, feb. 2017, doi: 10.1038/nature21056.

[27] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, e E. K. Oermann, «Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study», *PLOS Med.*, vol. 15, fasc. 11, p. e1002683, nov. 2018, doi: 10.1371/journal.pmed.1002683.

[28] Y. Shi *et al.*, «A Prospective, Molecular Epidemiology Study of EGFR Mutations in Asian Patients with Advanced Non–Small-Cell Lung Cancer of Adenocarcinoma Histology (PIONEER)», *J. Thorac. Oncol.*, vol. 9, fasc. 2, pp. 154–162, feb. 2014, doi: 10.1097/JTO.0000000000000033.