UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Triennale in Informatica

TESI DI LAUREA

Equità in Sistemi di Machine Learning: Stato dell'arte, Problemi e Sfide

RELATORE

Prof. **Fabio Palomba**

CORRELATORE

Dott. **Giammaria Giordano**

Università degli Studi di Salerno

CANDIDATO

Alessandra Parziale

Matricola: 0512108069

Anno Accademico 2021-2022

I computer sono incredibilmente veloci, accurati e stupidi.
Gli uomini sono incredibilmente lenti, inaccurati e intelligenti.
Insieme sono una potenza che supera l'immaginazione.
Albert Einstein

Sommario

I sistemi di machine learning sono sempre più presenti nella vita quotidiana delle persone. Spesso il loro utilizzo ha influenzato e screditato alcuni sottogruppi in base al genere, all'etnia o alla religione. Proprio per questo motivo il concetto di equità di tali sistemi ha succitato un particolare interesse recentemente. Questo elaborato è stato redatto allo scopo di esaminare proprio il concetto di equità, fornire una definizione di giudizio soggettivo e oggettivo generato sia dagli esseri umani sia dalle macchine e sulla base di questi concetti indagare a fondo sulle ingiustizie che possono provocare gli algoritmi dei sistemi intelligenti applicati in ambienti comuni e in ambienti sanitari. Durante la stesura sono stati identificati i dataset più utilizzati per tali problemi, le metriche per quantificare le ingiustizie, gli svantaggi e i vantaggi provocati dall'utilizzo di tali dispositivi intelligenti. Sulla base delle indagini svolte sono state inizialmente identificate le ingiustizie più frequenti con le relative cause e in un secondo momento sono state introdotte delle soluzioni valide per mitigare tali iniquità.

Indice

Elenco delle figure

Elenco delle tabelle

CAPITOLO 1

Introduzione

1.1 Motivazioni e obiettivi

L'evoluzione costante dei sistemi di machine learning ha causato una forte diffusione di tale tecnologia in ambienti comuni e abituali che fino a pochi anni fa non avevano nessun tipo di legame con la tecnologia intelligente.

Tale rapido sviluppo tecnologico ha introdotto molte importanti agevolazioni per il genere umano. Principalmente il miglioramento della qualità di vita del singolo individuo, in secondo luogo la possibilità di analizzare meticolosamente i dati, ottimizzare dei processi comuni oppure assistere i medici nelle azioni quotidiane.
Tuttavia ha anche rimarcato delle ingiustizie, infatti i giudizi soggettivi provocati da questi sistemi involontariamente, i pregiudizi e le discriminazioni [1][A21]che a loro volta generano considerevoli danni morali, sofferenze, ansie, turbamenti psicologici ed una perdita di fiducia nel sistema [30], sono tutti danni e problemi di iniquità che i sistemi intelligenti potrebbero causare.
In letteratura sono state individuate alcune applicazioni di sistemi intelligenti che hanno causato iniquità.
In ambienti ordinari i principali sistemi dove sono state individuate delle discriminazioni sono i consulenti intelligenti e i sistemi di autenticazione, successivamente i sistemi per la valutazione dei lavoratori e degli studenti, per il riconoscimento facciale, i Robo-advisor, i sistemi per i veicoli autonomi ed infine per i tribunali.
In ambienti sanitari invece sono state individuate delle discriminazioni nei sistemi intelligenti per assistere i medici durante la diagnosi delle malattie , nei sistemi dei dispositivi indossabili, nei sistemi delle cartelle cliniche elettroniche, nei sistemi per la comunicazione con il paziente, per l'analisi del tasso di recidiva e per l'adozione della chirurgia robotica.

Successivamente sono anche state individuate le principali cause legate alla penalizzazione di determinati gruppi di individui, nello sviluppo del sistema eseguito da persone comuni, il che comporta possibili errori umani e nell'utilizzo di dati inadeguati per addestrare gli algoritmi.

In questa tesi gli obiettivi principali prefissati sono stati: specificare il significato di equità e di pregiudizio, individuare i sistemi che presentano delle discriminazioni, analizzare i diversi tipi di ingiustizie, analizzare le cause e ricreare delle soluzioni valide. Per raggiungere questi obiettivi sono state eseguite più ricerche contemporaneamente su differenti database di articoli accademici, come per esempio Scopus, IEEE Xplore e ACM digital library.
Questo specifico approccio ha permesso di avere più informazioni sull'argomento di interesse ed eseguire un confronto approfondito in modo da rendere lo studio quanto più completo e preciso possibile.

1.2 Risultati ottenuti

In questa tesi di ricerca si è esteso lo stato dell'arte fornendo ulteriori informazioni sulle iniquità dei sistemi di machine learning. Precisamente sono state ricercate, studiate e riportate:

- Le metriche di equità più comuni.
- Le motivazioni che stimolano l'interesse dell'individuo per l'equità.
- Gli algoritmi intelligenti maggiormente propensi a discriminare.
- Le cause principali di discriminazioni.
- Le principali problematiche esterne dovute all'utilizzo di queste tecnologie.
- I vantaggi principali generati dall'utilizzo di queste tecnologie.

1.3 Struttura della tesi

I capitoli sviluppati in questa tesi di ricerca sono:

- **Capitolo 2: Stato dell'arte : Fairness in sistemi di intelligenza artificiale**, che analizza il significato di equità e di pregiudizio, ricerca le ragioni effettive che conducono una macchina ad eseguire un'azione svantaggiosa nei confronti di una determinata classe sociale e individua le materiche e i dataset più utilizzati in letteratura.
- **Capitolo 3: L'intelligenza artificiale nel settore sanitario**, che analizza alcuni algoritmi intelligenti in ambienti sanitari, ricerca eventuali discriminazioni, esamina i pregiudizi più comuni ed evidenzia le molteplici soluzioni per risolvere tali problematiche.
- **Capitolo 4: Metodologia**, che illustra il metodo adoperato per ultimare la ricerca, i motori di ricerca utilizzati, la query di ricerca, i risultati ottenuti prima dei filtri, filtri applicati, i criteri di inclusione ed esclusione ed infine i risultati finali ottenuti.
- **Capitolo 5: Analisi dei risultati**, che espone e analizza tutti i risultati ottenuti dalle ricerche.
- **Capitolo 6: Conclusioni**, che illustra una sintesi della ricerca e i risultati ottenuti.

CAPITOLO 2

Stato dell'arte : Fairness in sistemi di intelligenza artificiale

L'equità (Fairness) consiste nel trattare i soggetti in modo analogo indipendentemente dal sesso, dalla razza o dallo stato sociale

Comunemente il concetto di equità è oggi associato al giudizio in quanto esso consiste in una definizione, valutazione o critica di qualcuno oppure di qualcosa e pertanto può essere di due tipi, oggettivo o soggettivo.

Tipicamente agli esseri umani sono associati maggiormente i giudizi soggettivi in quanto l'uomo è facilmente condizionabile dalle emozioni, dalle esperienze vissute, dalla società o dalle opinioni altrui, contrariamente alle macchine sono associati esclusivamente i giudizi oggettivi dato che non possono essere influenzate da fattori esterni e per generare un output, quindi anche un giudizio, utilizzano esclusivamente determinati algoritmi[A32].

Considerando ciò, i giudizi umani risultano essere inevitabilmente soggettivi, svantaggiosi e difficili da modificare o eliminare, mentre i giudizi dalle macchine risultano essere più controllabili e meno discriminanti proprio perché sono oggetti e se venisse generata una discriminazione di qualsiasi tipo basterebbe modificare il suo algoritmo[A12][32].
Tuttavia sia che si tratti di giudizi soggettivi che di giudizi oggettivi, essi possono generare delle discriminazioni di nazionalità, etnia o religione in un determinato gruppo di persone.
Tali discriminazioni possono provocare, nella parte lesa, considerevoli danni morali in quanto potrebbero provocare non solo sofferenza e ansia ma anche un turbamento psicologico notevole.
Oltretutto questo ha comportato una completa perdita di fiducia nel sistema e un desiderio di equità da parte di chi è stato penalizzato stimolando di conseguenza un notevole incremento dell'interesse e dell'accettazione delle nuove tecnologie, considerate in grado di abbattere i pregiudizi sociali generati dagli esseri umani[30].

Queste nuove tecnologie si basano su macchine che non hanno una conoscenza incorporata ma utilizzano algoritmi forniti e impostati da persone fisiche. In un secondo momento l'algoritmo utilizzato inizierà ad apprendere automaticamente secondo le impostazioni fissate ed è proprio in questa fase che potrebbe essere originata una discriminazione di qualsiasi genere; infatti, se i dati iniziali sui quali l'algoritmo è stato addestrato non includono un numero sufficiente di membri di ogni sottogruppo esistente, l'output generato sarebbe penalizzante per uno specifico sottogruppo sul quale non possiede abbastanza informazioni[1] [A21].

Difatti in letteratura si associa l'origine del problema del pregiudizio all'uso inappropriato e alla qualità dei dati. Gran parte della letteratura che tratta della discriminazione causata dagli algoritmi si concentra sugli aspetti tecnici del pregiudizio e dell'equità[A5; A11].

Infatti si concentra principalmente sulla definizione delle diverse metriche per quantificare l'equità, queste sono:

- **Metriche di equità di gruppo:** essenzialmente confrontano il risultato dell'algoritmo di classificazione per due o più gruppi[A11].

- **Metriche basate sulla parità:** considerano i tassi positivi previsti tra diversi gruppi[A11].

- **Metriche basate su matrice di confusione:** tali metriche prendono in considerazione aspetti aggiuntivi come il tasso di veri positivi, il tasso di veri negativi , il tasso di falsi positivi e tasso di falsi negativi[A11].

- **Metriche di equità individuale:** non si concentrano sul confronto di due o più gruppi, ma considerano il risultato per ciascun individuo partecipante[A11].

Inoltre la letteratura con lo sviluppo della ricerca sull'equità, ha affrontato anche le teorie sul perché le persone si preoccupassero di tale argomento ed ha esaminato i processi di valutazione eseguiti che venivano utilizzati maggiormente[7].

Per quanto riguarda i motivi, questi sono stati classificati in tre gruppi:

- **Motivi strumentali**, secondo i quali le persone si preoccupano dell'equità perché gli concede il controllo sui propri risultati, cioè l'agire non per uno scopo immediato ma per un secondo fine o per un interesse non dichiarato, questa motivazione in letteratura è stata classificata come egoistica perché si concentra esclusivamente sul bisogno personale[7].

- **Motivi relazionali**, affrontano il concetto di come l'equità soddisfa il bisogno delle persone di sentirsi bene con sé stessi, cioè fornisce alle persone informazioni sulla loro relazioni con il gruppo che li circonda, anche questa motivazione si concentra sul bisogno del singolo e quindi in letteratura è considerata egoistica[7].

- **Motivi morali**, questa categoria, al contrario delle precedenti, si basa sull'importanza dei doveri e delle norme morali, cioè la capacità di operare in accordo ai principi universali e distinguere ciò che è giusto da ciò che è sbagliato, quindi affronta concetti riguardanti la società e non più i problemi del singolo individuo[7].

Per quanto riguarda le tecniche di valutazione, quelle che sono maggiormente affrontate sono:

- **Elaborazione controllata**, questo determinato tipo di approccio tratta l'equità come risultato di un processo basato sulla valutazione di regole o criteri di giustizia[7].

- **Elaborazione automatica**, gli individui attraverso questo approccio elaborano le informazioni utilizzando meno sforzo e meno risorse cognitive[7].

- **Teoria della giustificazione del sistema**, è una teoria sui giudizi di equità molto diffusa in letteratura, secondo la quale le persone hanno una tendenza a percepire i sistemi in modo equo. Di conseguenza, l'ingiustizia può essere percepita come particolarmente minacciosa e può portare le persone a utilizzare procedure sleali per preservare la loro convinzione di un sistema equo[7; 18].

Considernado che queste nuove tecnologie di apprendimento automatico sono utilizzate in contesti sempre più comuni dovrebbero anche garantire l'assenza di pregiudizi di genere, etnia o disabilità. Tuttavia, allo stato attuale, possiamo ancora trovare esempi celebri di discriminazione provocata da sistemi intelligenti:

Facebook è stata più volte al centro di polemiche sul razzismo e discriminazioni[37].

Per la prima volta nel 2016 i sistema di distribuzione degli annunci di lavoro di Facebook, basato su algoritmi e tecniche di intelligenza artificiale, sono stati accusati di escludere le donne da alcune opportunità lavorative anche se le tipologie di lavoro richiederebbero le stesse qualifiche e capacità indistintamente per uomini e per donne.

Una seconda volta nel 2017, Facebook è stata accusata di razzismo per aver introdotto nel suo servizio di traduzione automatica una discriminazione di razza, poiché l'algoritmo alla base del sistema tendeva ad associare ingiustamente alcune parole razziste o discriminatorie ad un gruppo specifico di soggetti[9].

Un secondo esempio di discriminazione è Deliveroo, una compagnia di consegna di cibo, e il suo algoritmo per la scelta e il giudizio dei riders. Infatti tale algoritmo è stato condannato perché discriminava i riders autonomi, cioè non assunti dall'azienda[2].

Altri esempi celebri sono quelli di Amazon, Google e Microsoft.
Infatti sono stati testai i diversi servizi di Google Cloud Vision, Microsoft Azure Computer Vision e Amazon Rekognition, con foto di politici, donne e uomini. Gli algoritmi etichettavano le donne solo con caratteristiche estetiche come "giovane" o "bella" a differenza degli uomini che venivano etichettati per le loro qualità intellettuali[34]

In Germania è stato evidenziato un problema simile provocato dall'algoritmo utilizzato da Google per associare determinate parole alle immagini. Tale algoritmo era poco cortese con il genere femminile rispetto al genere maschile, infatti associava alle ricerche di genere femminile solo immagini riguardanti esclusivamente l'estetica, la sensualità o la forma fisica mentre agli uomini sono associate immagini riguardanti l'ambito lavorativo o professionale[5].

Inoltre, per quanto riguarda Amazon, è stato al centro anche di una seconda polemica nel 2014 sviluppando un sistema intelligente capace di valutare i candidati per un lavoro, con l'obiettivo di meccanizzare la ricerca e selezionare i "migliori" assegnando un giudizio da una a cinque stelle.
Tuttavia nel 2015 l'azienda si è resa conto che il sistema non considerava tutti i candidati in modo equo per lavori di sviluppo software o incarichi tecnici. Questo perché gli algoritmi sono stati addestrati osservando dati raccolti in un periodo nel quale nell'industria tecnologica vi era un predomino del genere maschile [12].

Per risolvere i problemi provocati si potrebbero analizzare e correggere i dati raccolti, inserendo più informazioni o rappresentando meglio gruppi sfavoriti, piuttosto che cercare di correggere l'elemento discriminatorio all'interno del modello [11].

2.1 Definizione di Fairness nel contesto IoT

Internet of things (IoT) è un sistema di dispositivi interconnessi con la capacità di monitorare e trasferire dati.

Una rete di oggetti fisici incorporati con elettronica, software o sensori, capace di controllare oggetti che permettono l'integrazione diretta tra il mondo fisico e i sistemi basati su computer.

L'IoT ha quindi introdotto l'automazione in sistemi che utilizziamo giornalmente, come quelli per la gestione sanitaria, un settore particolarmente attivo nell'adozione di queste nuove tecnologie.

Nel quale, la registrazione dei dati sensibili, la diagnosi intelligente o l'andamento di malattie sono tutte attività gestita da questi dispositivi intelligenti ed un qualsiasi malfunzionamento può provocare gravi conseguenze per un paziente.

L'utilizzo di queste tecnologie informatiche a vantaggio della salute umana è denominato *e-Health*. Esistono quattro livelli nell'architettura *e-Health* basata su IoT: livello di interfaccia, di servizio, di rete e di rilevamento.

Nel **livello di rilevamento** vengono raccolte le informazioni sullo stato di salute dei pazienti, attraverso il **livello di rete** sono inviate per l'elaborazione, nel **livello di servizio** sono analizzate per determinare lo stato di salute del paziente e infine i risultati sono comunicati a medici e pazienti tramite il **livello di interfaccia**[A14].

I vantaggi dell'*e-Health* sono i costi ridotti e i minimi tempi di attesa, gli svantaggi provocati invece riguardano la poca affidabilità dei sistemi di sicurezza, infatti un possibile attacco al sistema di prescrizione, agli orari delle visite ospedaliere e alle ambulanze intelligenti potrebbe essere esternamente pericoloso.

Purtroppo i sistemi IoT non sono ottimali in circostanze di privacy e sicurezza per diversi motivi, come ad esempio l'eccessiva dinamicità che provoca il loro cambiamento continuo, oppure per la loro eterogeneità rispetto ai mezzi di comunicazione, piattaforme e dispositivi, o per la possibilità di includere anche sistemi che non sono stati progettati per essere connessi ad internet.

Tuttavia la OWASP INTERNET OF THINGS PROJECT, progettata per aiutare produttori, sviluppatori e consumatori a comprendere meglio i problemi di sicurezza associati all'IoT, ha identificato la vulnerabilità più grande di questi sistemi nella mancanza di tecniche di sicurezza note, come autenticazione, crittografia o controllo dell'accesso.

Quindi nonostante le tecnologie IoT introducano nuove opportunità è comunque fondamentale e necessario trovare una soluzione per garantire sicurezza e privacy a tutti coloro che usufruiscono di questi sistemi [A18; A22] [8].

2.2 Problemi Di Fairness in Sistemi di AI

Una prima applicazione dell'intelligenza artificiale *(Artificial Intelligence (AI))* al quotidiano è stata atraverso l'uso dei AI-ADVISOR per migliorare il processo decisionale umano.
Alcuni esempi sono SIRI di Apple o ALEXA di Amazon che aiutano gli utenti a decidere in base alle loro preferenze personali, oppure sistemi di supporto per decisioni cliniche progettate per assistere gli operatori sanitari nel processo decisionale o anche i ROBO-ADVISOR che forniscono assistenza basata su algoritmi per la pianificazione finanziaria.
Tuttavia dato che in molti casi l'intelligenza artificiale viene associata al raggiungimento della **"superIntelligenza"** di una macchina, cioè la capacita di superare le prestazioni degli esseri umani, si è proposto di creare sistemi intelligenti in grado di monitorare costantemente gli stati mentali o l'ambiente circostante e sulla base di questi, suggerire la migliore linea d'azione in una determinata situazione[A7; A9].

A tale scopo è stata introdotta la *Brain Computer Interface (BCI)* che si occupa di realizzare un canale di comunicazione diretto tra il cervello e un dispositivo esterno.
Ne esistono di due tipi, *BCI invasivi*, che si basano sull'utilizzo di impianti applicati attraverso un intervento chirurgico, considerati più rischiosi ma in grado di fornire risoluzioni più accurate, oppure *BCI non-invasivi* che permettono di catturare l'attività celebrare attraverso altre tecniche come l'elettroencefalografia. È stato possibile applicare il *BCI* nell'ambito l'*AI* per migliorare le capacità di attenzione, regolare gli stati emotivi, moderare le condizioni ambientali, massimizzare l'efficienza di un lavoratore, oppure utilizzarlo nell'ambito delle attività condotte dell'Esercito, infatti sia le forze armate statunitensi sia quelle cinesi hanno valutato la possibilità di utilizzarlo per consentire il trasferimento diretto di pensieri da un cervello all'altro[A7; A9].

Purtroppo con l'utilizzo dei sistemi intelligenti, la privacy del singolo individuo non è del tutto rispettata, di conseguenza molte persone essendo già restie ad utilizzare tecnologie più semplici, come SIRI di Apple o ALEXA di Amazon, per paura che la propria privacy personale possa essere violata, sarebbero ancora meno propense ad accettare una tecnologia così avanzata in grado di accedere a questioni ancora più personali e capaci di alterare decisioni morali o sanitarie.
Inoltre con l'introduzione di queste tecnologie si rischierebbe di amplificare il divario digitale già esistente e non solo generare pregiudizi di tipo economico, cioè tra coloro che non potrebbero permettersi tale tecnologia e quindi essere in "svantaggio" rispetto ad altri, ma anche penalizzare ingiustamente lavoratori o studenti per dei bassi punteggi di attenzione, nel caso in cui la tecnologia venga utilizzata per monitorare e aumentare l'attenzione a scuola o a lavoro.[A6]
Tale discriminazione si potrebbe riscontrare anche in sistemi più "basici" quali i consulenti *AI*, che potrebbero anche mostrare pregiudizi nelle loro raccomandazioni andando a discriminare alcuni gruppi sociali, sia a causa della procedura utilizzata sia a causa dei dati su cui farebbero affidamento[A7; A9].
Una possibile soluzioni, proposta da molti[A10], sarebbe quella di garantire che gli algoritmi siano addestrati su dataset contenenti dati riguardanti ogni sottogruppo esistente in modo da non generare divari significativi di etnia o genere.

2.3 DataSet esistenti

Di seguito è riportata una panoramica dei dataset utilizzati per affrontare problemi di non equità (FAIRNESS) negli algoritmi di apprendimento.

- **Adult Dataset**
  Adult dataset [1] è utilizzato per gli studi di classificazione con equità[14; 18], questo ha il compito di classificare e prevedere, in base a delle caratteristiche, se il reddito annuo di una persona supera i 50.000 dollari.

  Tale dataset è composto da 41 attributi come età, istruzione dell'individuo, genere o ore di lavoro, è dominato da istanze maschili, popolazione caucasica e da un'età compresa tra i 25 e i 60 anni[A16].

- **KDD Census-Income Dataset**
  KDD Census-Income dataset [2] è stato raccolto e assemblato dal Current Population Surveys effettuato dall'U.S. Census Bureau nel 1994.

  La sua attività di previsione è la medesima dell'Adult dataset con la differenza che il campo obiettivo è stato ricavato del campo del reddito totale della persona piuttosto che dal reddito lordo[19; 28].

  Tuttavia anche in questo specifico caso il dataset è sbilanciato poiché è dominato da istanze maschili e appartenenti alla popolazione caucasica. [A16].

- **German Credit Dataset**
  German Credit dataset [3] tedesco formato da campioni titolari di conti bancari, utilizzato per gli studi di apprendimento con equità[17; 35].

  Viene utilizzato per la valutazione del rischio, quindi per analizzare e determinare se è rischioso concedere o meno del credito.

  I "clienti" sono classificati in "buoni" o "cattivi" e contiene 21 attributi totali come età, genere, importo del credito o occupazione.

  Anche in questo caso il dataset è dominato da istanze maschili e età superiore ai 25 anni[A16].

[1] http://archive.ics.uci.edu/ml/datasets/Adult
[2] https://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD)
[3] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

- **Dutch Census Dataset**
  Dutch Census dataset [4] rappresenta gruppi di persone nei Paesi Bassi dell'anno 2001, è stato utilizzato per classificare l'occupazione delle persone in base al tipo di professione (di alto livello o di basso livello)[13; 21].

  Tale dataset è composto da 12 attributi ed è stato pre-elaborato scartando i campioni non adeguati come i minorenni oppure la cui professione è sconosciuta o di livello medio[A16].

- **Bank Marketing Dataset**
  Bank Marketing dataset [5] è di marketing bancario, precisamente è formato da campagne di marketing di un istituto bancario portoghese tra il 2008 e il 2013[22; 38].

  È stato utilizzato per prevedere se un determinato cliente potrebbe effettuare o meno una sottoscrizione di certificati di deposito ed ha un totale di 17 attributi come età, lavoro, stato coniugale o formazione scolastica[A16].

- **COMPAS Dataset (Correctional Offender Management Profiling for Alternative Sanctions)**
  COMPAS dataset [6] è stato utilizzato nei tribunali degli Stati Uniti per prevedere il rischio di recidiva di un criminale, cioè per analizzare se un criminle viene nuovamente arrestato entro due anni dal primo arresto[6; 16].

  Tuttavia è stata riscontrata una discriminazione di razza, poiché il numero di criminali di colore arrestati è di gran lunga superiore rispetto a quello dei criminali bianchi nonostante abbiano gli stessi punteggi[A16].

- **Ricci Dataset**
  Ricci dataset [7] è generato dal caso Ricci v. DeStefano riguardante il diritto al lavoro negli Stati Uniti, è utilizzato per prevedere se un individuo ottiene una promozione in base al risultato di un esame sostenuto[20; 33].

  Si tratta di un dataste abbastanza piccolo, infatti contiene solo 6 attributi, la percentuale di promozioni concesse a persone che hanno superato l'esame è il 100 tuttavia vi è anche in questo caso una discriminazione di razza in quanto il numero di promozioni concesse a persone appartenenti alla popolazione caucasica è di gran lunga superiore rispetto ad altre[A16].

[4] https://archive.ics.uci.edu/ml/datasets/census+income
[5] https://archive.ics.uci.edu/ml/datasets/bank+marketing
[6] https://www.kaggle.com/danofer/compass
[7] https://www.key2stats.com/data-set/view/690

- **OULAD Dataset (Open University Learning Analytics)**
  OULAD dataset [8] possiede alcuni dati raccolti in Inghilterra dall'Open University (OU) nel 2013, contiene informazioni circa gli studenti e il loro apprendimento virtuale.
  Questo dataset è utilizzato per prevedere il successo degli studenti con questa tipologia di apprendimento[24; 27].

  Questo specifico dataset è spesso utilizzato per studi su problemi di fairness, contiene 12 attributi ed è stato riscontrato che la percentuale di studenti maschi con un'istruzione è di gran lunga superiore rispetto a quella di studentesse femmine[A16].

[8] https://archive.ics.uci.edu/ml/datasets/Open+University+Learning+Analytics+dataset

CAPITOLO 3

L'intelligenza artificiale nel settore sanitario

L'intelligenza artificiale e l'apprendimento automatico sono tecnologie ampiamente utilizzate da aziende, governi o organizzazioni di ogni genere per migliorare la qualità operativa e assistere nel miglior modo possibile il processo decisionale.

Un settore particolarmente interessato all'utilizzo di queste nuove tecnologie è quello medico, poiché l'utilizzo di tecniche di machine learning è in grado di fornire strumenti all'avanguardia agli operatori sanitari e una personalizzazione delle cure dei pazienti. Infatti un sistema intelligente è in grado di comprendere, elaborare e interpretare i dati raccogliendo ogni tipo di feedback, per migliorarsi continuamente. Oltretutto tale sistema riesce anche di interpretare il linguaggio medico facilitando la collaborazione con medici e pazienti, quindi l'adozione di queste nuove tecnologie può generare la totale automazione e l'agevolazione del lavoro degli specialisti[29] [A13].

Sono molteplici le aziende che si sono interessate all'applicazione dell'intelligenza artificiale e dell'apprendimento automatico al settore medico.

Un esempio è Google con il progetto *DeepMind Health* basato sullo sviluppo di un sistema in grado di processare una quantità smisurata di informazioni mediche in pochi minuti.
In questo modo i processi sanitari di qualsiasi genere verrebbero velocizzati di molto, inoltre i ricercatori coinvolti in tale progetto hanno anche cercato di elaborare modelli capaci di immaginare le conseguenze di una determinata azione prima di intraprenderla, in modo da evitare possibili complicazioni[31].

Un secondo esempio proposto è *Verily*, un ramo di Google che si occupa della scienza della vita, il quale ha sviluppato un progetto chiamato *Baseline Study* per la raccolta e la gestione dei dati genetici, con lo scopo di adottare alcuni degli algoritmi di Google per studiare e analizzare

ciò che compromette o meno la salute di una persone. Per questo progetto sono state utilizzate delle tecnologie altamente all'avanguardia per il monitoraggio di alcune malattie, come ad esempio l'utilizzo delle lenti a contatto intelligenti per misurare il livello di zucchero nel sangue[31].

Molto attiva in ambito medico è anche l'*IBM(International Business Machines Corporation)* che ha sviluppato l'intelligenza artificiale *Watson* in grado di anticipare di due anni, rispetto ai metodi tradizionali, la diagnosi di insufficienza cardiaca[31].

Anche Intel ha esplorato recentemente il settore medico, in particolare esso si è concentrato sui possibili tumori ai polmoni, infatti insieme ad altri partner ha sviluppato un algoritmo capace di leggere le radiografie e i dati medici di un determinato paziente per anticipare la diagnosi e seguire l'avanzamento del tumore [31].

3.1 Applicazione di sistemi di intelligenza artificiale in medicina

Le tecniche di machine learning utilizzate in ambito medico sono in grado di agevolare sia il paziente, individuando più velocemente le patologie, fornendo una diagnosi accurata e specifica tenendo conto delle esigenze del singolo e confrontando la sua storia clinica con la collezione di dati raccolti da milioni di altri pazienti, sia il medico in quanto permette lo snellimento di alcuni passaggi burocratici e amministrativi che risultano inefficienti e macchinosi, permette inoltre la focalizzazione immediata dei dati più importanti di un paziente utilizzando dei motori di ricerca che evidenziano le informazioni nella cartella clinica e dà la possibilità di raggiungere un maggior numero di specialisti in nodo da fornire una valutazione medica senza un diretto coinvolgimento.

Inoltre l'utilizzo di tecniche di intelligenza artificiale ha recato un miglioramento e una totale automazione di alcune fasi cliniche fondamentali, come **la prognosi, la diagnosi e il trattamento**.

La prognosi è un procedimento utilizzato per predire lo sviluppo di una certa malattia. Un modello di machine learning è in grado di agevolare questa predizione in quanto permette ai medici di prevedere eventi futuri come tra quanto un paziente potrà tornare alla vita di tutti i giorni, oppure quanto velocemente progredirà una malattia. Tale modello dovrà avere a disposizione un quadro completo del paziente, inclusi i risultati delle terapie alle quali si e sottoposto, i risultati di test fenotipici o immagini mediche.

La diagnosi è un processo per determinare quale malattia o condizione genera determinati sintomi in un paziente. Uno studio fatto dall'*American Institute of Medicine* ha evidenziato che almeno una volta nella vita un paziente riceve una diagnosi errata. È possibile ridurre al minimo questo tipo di errore attraverso le tecniche di intelligenza artificiale le quali, con i dati collezionati durante terapie, possono individuare le diagnosi più probabili durante una visita clinica e prospettare quali condizioni si manifesteranno in futuro nel paziente.

Il trattamento è l'attuazione concreta di quanto necessario per poter portare alla guarigione di un paziente, potrebbero però presentarsi delle variazioni circa le modalità di trattamento di certi sintomi. Un algoritmo di machine learning può aiutare i dottori a identificare un trattamento da preferire rispetto ad un altro, eseguendo anche un confronto con ciò che il dottore prescriverebbe a un paziente e con un trattamento suggerito dal modello.

Numerosi algoritmi di machine learning e di intelligenza artificiale, approvati dalla *Food and Drug Administration (FDA)* l'ente governativo statunitense che si occupa della regolamentazione dei prodotti farmaceutici, sono utilizzati in ambito medico.

Uno di questi è l'algoritmo *Naïve Bayesian (NB)*, basato su particolare modello probabilistico che identifica l'incertezza di un modello attraverso la probabilità. È utilizzato in diversi sistemi in ambito medico poiché richiede pochi dati per l'addestramento, tuttavia non è uno degli algoritmi più accurati [25].

Un secondo esempio è *Artificial Neural Network (ANN)(rete neurale artificiale)*, cioè un modello composto da neuroni artificiali, ispirato dalla semplificazione di una rete neurale biologica. Tale rete in medicina è utilizzata per la diagnosi, per identificare patologie o per la valutazione di un quadro clinico. Il vantaggio di queste reti neurali è che sono flessibili poiché possono essere aggiornate per adattarle ad ambienti molto dinamici come quelli medici [15; 25].

Un terzo esempio è il *Support Vector Machines (SVM)*, il quale viene utilizzato per classificare i soggetti all'interno di gruppi inoltre è considerato uno dei classificatori che genera risultati più accurati. Infatti in uno studio l'*SVM* è stato utilizzato per classificare e rilevare le varie tipologie di caduta negli anziani, durante la ricerca sono stati utilizzati cinque classificatori, ma *SVM* ha raggiunto i risultati più precisi. Tuttavia lo svantaggio di questa tecnica è la sua mancanza di trasparenza dei dati [25; 29][A3].

Attualmente questi algoritmi sono applicati a diversi rami della medicina:

- **Cardiologia, precisamente alla fibrillazione atriale e al rischio cardiovascolare**
  L'azienda di dispositivi medici *AliveCor*, si è particolarmente interessata alla diagnosi precoce della **fibrillazione atriale**, la quale è stata anche una delle prime applicazioni di intelligenza artificiala in ambito medico, l'azienda in seguito all'approvazione della FDA ha sviluppato una sua applicazione mobile, Kardia, che consente il monitoraggio ECG e il rilevamento della fibrillazione atriale.

  Anche la Apple, dopo aver ricevuto l'approvazione della FDA, ha inserito nel suo *Apple Watch 4* il monitoraggio dell'ECG e il rilevamento della **fibrillazione atriale** con la possibilità di condividere le informazioni con il proprio medico curante, tramite uno smartphone.
  L'utilizzo di questa applicazione concede, a chi lo indossa, maggiori probabilità di identificare la fibrillazione rispetto ai semplici controlli eseguiti di routine. Nonostante questo, è una

tecnologia molto criticata per la percentuale alta di falsi positivi o per l'impossibilità di essere utilizzata dai pazienti più anziani.

Per quanto riguarda il **rischio cardiovascolare**, l'intelligenza artificiale è stata molto spesso utilizzata per prevedere più accuratamente e più velocemente, rispetto alle normali tecniche tradizionali, il rischio di malattie come la sindrome coronarica acuta o l'insufficienza cardiaca, considerate ormai le principali cause di morte nel mondo.
Per questo motivo i dispositivi informatici indossabili utilizzati per supervisionare le possibili malattie cardiovascolari hanno suscitato l'interesse di molti. Infatti il laboratorio di Neuronica del politecnico di Torino ha sviluppato il Vital-ECG [26] un dispositivo indossabile a basso costo, facile e comodo da usare per le piccole dimensioni.
Tale dispositivo è stato sviluppato con un algoritmo di apprendimento automatico attraverso il paradigma internet of things (IoT) per dispositivi sanitari smart ed è in grado di tracciare l'elettrocardiogramma, la saturazione dell'ossigeno, l'attività fisica del paziente o la pressione arteriosa e di elaborare tali dati per generare un allarme sullo stato di salute del paziente tramite l'applicazione a esso abbinata. Tuttavia sono stati sviluppati molteplici dispositivi per controllare l'attività cardiaca, come il QardioCore una cintura che avvolge il torace controlla l'attività cardiaca, oppure l'HeartGuide un vero e proprio orologio in grado di monitorare la pressione arteriosa. [A3; A15].

- **Medicina polmonare**
  Recentemente è stato evidenziato che un software basato sull'intelligenza artificiale riesce a fornire un'interpretazione più accurata e può servire come uno strumento di supporto per le decisioni riguardanti l'interpretazione dei possibili risultati dei test di **funzionalità polmonare**. Anche in questo caso la tecnologia sviluppata ha ricevuto diverse critiche, una delle quali ha evidenziato che il tasso di diagnosi accurata era inferiore alla media [A3].
- **Endocrinologia**
  Un monitoraggio continuo della glicemia consente ai pazienti diabetici di avere una quantità maggiore di informazioni sull'eventuale **variazione dei livelli di glucosio nel sangue**.

  A questo ramo della medicina si è particolarmente interessata la Medtronic, la quale dopo aver ricevuto l'approvazione della FDA ha sviluppato il sistema, *Guardian*, che consente il monitoraggio della glicemia abbinato all'utilizzo dello smartphone.

  Oltretutto l'azienda ha anche utilizzato *Watson*, un sistema intelligente sviluppato da *IBM*, per il sistema *Sugar.IQ*, "un'assistente diabetico", per aiutare i clienti a monitorare gli episodi di ipoglicemia con misurazioni ripetute [A3].
- **Gastroenterologia**
  Questo specifico ramo beneficia di un'ampia gamma di applicazioni di intelligenza artificiale. I gastroenterologi hanno utilizzato le reti neurali artificiali per diagnosticare **il reflusso gastroesofageo e la gastrite atrofica** oppure per predire le ragioni del **sanguinamento gastrointestinale**, dell'evoluzione del cancro esofageo e delle malattie intestinali [A3].

- **Neurologia, precisamente all'epilessia, alla valutazione dell'andatura, della postura e alla diagnosi del cancro**
  L'azienda, Empatica, dopo aver ricevuto l'approvazione della FDA si è concentrata nello sviluppo di un'applicazione mobile in grado di rilevare una **crisi epilettica** e di allertare, con la posizione del paziente, il medico di riferimento e i parenti.

  Al contrario dei dispositivi per il monitoraggio cardiaco, in questo caso specifico non vi sono barriere nell'adozione di questi dispositivi da parte dei pazienti.

  Per quanto riguarda **l'andatura e la postura**, sono stati studiati e sviluppati dei sensori indossabili dai pazienti con sclerosi multipla o morbo di Parkinson, che consentono una verifica dell'andamento della malattia.

  Inoltre è stato anche sviluppato un tipo di algoritmo basato sull'intelligenza artificiale in grado di diagnosticare il cancro al cervello con grande precisione, consentendo al medico di guadagnare tempo prezioso [A3].
- **Fisioterapia**
  Attraverso l'intelligenza artificiale sono state sviluppate tecnologie anche nel campo della **fisioterapia**.
  Queste si basano sul principio di collaborazione infatti mentre il terapista umano guida il paziente nel **compiere determinati movimenti** per il probabile recupero di una specifica funzionalità motoria, l'intelligenza artificiale è in grado di riconoscere immediatamente se gli esercizi svolti dal paziente sono corretti e se porteranno ad un risultato effettivo.

  Il loro compito, infatti è quello di osservare e giudicare costantemente il paziente e, qualora non dovese riuscire a compiere il movimento previsto, aiutarlo con la quantità di forza necessaria

  L'*Universidad Carlos III* di Madrid e l'*Hospital Universitario Virgen del Rocío* di Sevilla hanno creato un sistema con l'utilizzo dell'intelligenza artificiale e della robotica per aumentare l'interazione con il paziente.
  Tale progetto si basa su robot umanoidi in grado di parlare, emettere suoni e luci per aumentare l'attenzione del paziente.
  Attraverso uno studio è stato dimostrato che questo tipo di attività abbia aumentato il grado di mobilità degli altri superiori di bambini colpiti da paralisi celebrale [4; 23] [A3].
- **Diagnostica predittiva**
  La diagnostica predittiva attraverso l'utilizzo e l'interpretazione dei dati permette di cogliere i primi segnali di alcune malattie. Sono state introdotte tecniche intelligenti anche in questo settore per aiutare i medici a fare diagnosi più accurate, ridurre gli errori e sviluppare metodi per il trattamento individualizzato.

Anche l'ospedale universitario di *Harvard*, il *Beth Israel Deaconess Medical Center*, utilizza l'intelligenza artificiale per diagnosticare **malattie del sangue potenzialmente mortali** in una fase molto precoce.

Precisamente attraverso l'utilizzo di particolari microscopi dotati di intelligenza artificiale hanno ottimizzato la ricerca di batteri nocivi in campioni di sangue. Gli scienziati hanno usato 25.000 immagini di campioni di sangue per insegnare alle macchine come cercare i batteri e le macchine hanno poi imparato a identificare e prevedere i batteri nocivi con una precisione del 95%[3].

- **Medicina di precisione**
  La medicina di precisione basata su tecniche di machine learning permette di sviluppare **modelli predittivi personalizzati**, con la possibilità di personalizzare le cure anziché usare un approccio standard.

  Alla *Harvard Medical School* hanno sperimentato il *Buoy Health*, un controllore virtuale dei sintomi e delle cure. Il suo funzionamento, basato su un chatbot, cioè un software che simula ed elabora le conversazioni umane, ascolta i sintomi di un paziente, le sue preoccupazioni per la salute e guida nel percorso di cura in base alla diagnosi [3].

3.2 Fairness nel settore medico

Considerando che le decisioni in ambito medico sono sempre più influenzate dalla presenza dell'intelligenza artificiale è necessario considerare anche tutti i problemi di equità che potrebbero generarsi.
Purtroppo l'assistenza e l'intelligenza sanitaria sono da molto tempo al centro di polemiche basate su pregiudizi e discriminazioni, uno dei primi esempi risale agli anni 70 quando un algoritmo utilizzato dalla st george's hospital medical school di Londra discriminava i candidati per la facoltà di medicina in base al genere o all'etnia.[A1].
Infatti l'apprendimento automatico in campo sanitario utilizza da sempre dati generati da medici e pazienti per addestrare gli algoritmi e quindi le macchine a svolgere compiti specifici con capacità sovrumane.
Talvolta tali dati includono anche delle disparità e se vengono utilizzati per sviluppare un modello di intelligenza artificiale potrebbe amplificarsi il pregiudizio e il modello potrebbe anche prendere decisioni sfavorevoli nei confronti di un particolare gruppo di persone in base all'età, al sesso, all'etnia o al livello economico.[A11].
Ovviamente tale pregiudizio potrebbe danneggiare la qualità del servizio sanitario offerto.
L'*American Medical Association* ha rilasciato delle raccomandazioni per promuovere lo sviluppo di un'intelligenza di alta qualità che identifichi e adotti delle misure per affrontare le disparità sanitarie e sostenere la progettazione di sistemi di apprendimento automatico che promuovano l'equità sanitaria.

Purtroppo l'assistenza e l'intelligenza sanitaria sono da molto tempo al centro di polemiche basate su pregiudizi e discriminazioni.

Un primo esempio in cui l'apprendimento automatico potrebbe danneggiare un gruppo è quello del sistema di monitoraggio, ad esempio per aiutare una squadra di medici a **identificare i pazienti ricoverati ad alto rischio di deterioramento che richiedono il trasferimento immediato** in un'unità di terapia intensiva.

Per costruire il modello sono state raccolte e analizzate le registrazioni di alcuni pazienti che avevano un deterioramento clinico e quelli che non lo avevano.
Il modello esamina i rischi e indica se è necessario un trasferimento immediato.
Tuttavia, se nei dati di addestramento sono stati inclusi pochi pazienti di un determinato gruppo, il modello potrebbe essere impreciso per loro e se i team clinici iniziassero a fare affidamento esclusivamente su tale predizione, per identificare i pazienti a rischio, senza considerare che il sistema potrebbe non rilevare i pazienti da trasferire di un determinato gruppo, potrebbero scartare e danneggiare in modo significativo tali pazienti.

Un secondo esempio è un sistema in grado di **monitorare la durata di ricovero**, infatti se un ospedale si affidasse ad un modello basato su variabili cliniche e sociali per prevedere quali pazienti ricoverati potrebbero essere dimessi prima, in modo da poter identificare e indirizzare le risorse

per la gestione di specifici casi, potrebbe identificare una maggiore durata di ricovero nei pazienti provenienti da quartieri economicamente svantaggiati o prevalentemente afroamericani.

In questo modo il modello potrebbe allocare in modo sproporzionato risorse per gestire solo i casi dei pazienti provenienti da quartieri più ricchi o prevalentemente bianchi e discriminare gli afroamericani o i quartieri più poveri.

Quindi data la probabilità alta di discriminazioni nei modelli di apprendimento automatico è fondamentale mitigare la distorsione dei dati, coinvolgere tutte le popolazioni emarginate, attuare approcci tecnici diversi per l'addestramento dei modelli di apprendimento automatico.
Tuttavia la necessità di equità nell'assistenza sanitaria non si limita alla progettazione di algoritmi e quindi alla raccolta dei dati ma anche ad altri aspetti come la loro vera e propria progettazione del dispositivo, infatti il *QardioCore*[1] generava una discriminazione di genere in quanto, dovendo essere posizionato sul torace, non monitorava correttamente le donne per la presenza del seno, oppure anche l'*HeartGuide*[2] dato che si basava sul principio di rigonfiamento e sgonfiamento del canturino discriminava il genere femminile poiché risultava essere eccessivamente grande per le donne[A15].
Evidenziando questi aspetti sono state considerate molteplici soluzione per risolvere i pregiudizi causati, una **prima soluzione** è stata proposta dal programma statunitense *All of US* che mira a **raccogliere tutti i dati possibili da tipi diversi di persone**, in modo da avere a disposizione più informazioni anche dei gruppi meno considerati.

Una **seconda soluzione** è stata proposta dall'*IBM* con lo sviluppo di *AI Fairness 360*, un tipo di **toolkit online**, per aiutare ad esaminare, segnalare e mitigare le discriminazioni e i pregiudizi nei sistemi di apprendimento automatico.

Una **terza soluzione** è stata **l'identificazione di tre caratteristiche fondamentali** che i sistemi di apprendimento automatico, applicati all'ambiente sanitario, dovrebbero rispettare per i principi di uguaglianza ed equità, che sono :

La parità per i pazienti, cioè la garanzia che tutti i gruppi presi in considerazione traggono gli stessi benefici dall'implementazione di modelli di apprendimento automatico.

La parità di prestazioni, cioè la garanzia che un modello è ugualmente accurato per tutti i pazienti.

L'allocazione equa, cioè la garanzia che le risorse siano allocate proporzionalmente a tutti i pazienti.

Tuttavia risolvere questo tipo di problemi in abito medico risulta più complesso rispetto ad altri tipi di ambiti in quanto richiede non solo l'intervento della comunità informatica ma anche la collaborazione di medici e pazienti [A2; A8; A17; A19].

[1] Citato del paragrafo 3.1
[2] Citato del paragrafo 3.1

3.3 Problematiche provocate dai sistemi di intelligenza artificiale in medicina

L'ambito sanitario con l'applicazione dell'intelligenza artificiale genera non solo problemi di equità ma anche di diversa natura causati da fattori esterni.

Alcuni di questi problemi sono:

- **Fattori ambientali e i flussi di lavoro**
  Le prestazioni e l'efficacia di un modello intelligente possono essere influenzate sia dai **fattori ambientali** che dai **flussi di lavoro**.

  Un particolare studio condotto da Google Health [36] ha valutato le prestazioni di un sistema di intelligenza artificiale per lo screening della retinopatia diabetica, una complicazione del diabete che colpisce gli occhi, tale sistema è stato sviluppato per identificare e valutare la malattia in minor tempo rispetto alla prassi standard.

  Il sistema ha riscontrato dei problemi quando veniva applicato alle cliniche thailandesi poiché le condizioni e il flusso di lavoro in queste specifiche cliniche ha compromesso la qualità delle immagini, rendendole non adeguante agli standard elevati del sistema.
  Inoltre l'instabilità della connessine internet ha limitato la velocità di elaborazione causando dei tempi di attesa molto più lunghi.

  Lo studio ha quindi evidenziato l'importanza di convalidare i sistemi di intelligenza artificiale in **ambienti di diversa natura** e di considerare i feedback degli utenti per l'apprendimento e il miglioramento del sistema.

  Tuttavia nel settore sanitario risulta molto dispendioso, in termini di tempo e di costo, raccogliere tutti i feedback necessari, poiché potrebbe volerci molto tempo per valutare l'effetto di una terapia e gli esiti prodotti.

  Un altro modo potrebbe essere l'utilizzo di dati sintetici o sfruttare un ambiente simulato [A8].

- **Privacy e sicurezza**
  I dati sono fondamentali per un sistema basato sull'intelligenza artificiale, tuttavia questi dati per legge sono considerati **sensibili** e quindi di scarsa reperibilità generando in questo modo l'impossibilità di produrre nuovi modelli di machine learning, **una soluzione** è quella di ottenere il consenso da parte del paziente interessato.
  Tuttavia la continua raccolta e analisi dei dati dei pazienti genera preoccupazione negli stessi, infatti questi sono sempre più restii alla condivisione dei dati personali per la propria **privacy e sicurezza**, causando, anche in questo caso, un problema per la creazione dei nuovi sistemi intelligenti.

Sarebbe quindi l'ideale, per la creazione e l'adozione di queste nuove tecnologie, avere a disposizione **dei dataset costruiti su cartelle cliniche reali**, tuttavia è comunque essenziale avere un sistema di **archiviazione dei dati sicuro e ben controllato** per garantire la sicurezza e il rispetto della privacy dei pazienti[29] [A8].

- **Disumanizzazione**
  La tecnologia dell'*intelligenza clinica ambientale (ACI)* è spesso considerata sensibile perché comprende sia i pazienti sia i medici ed è in grado di studiare e analizzare la visita di un paziente e compilare automaticamente la cartella clinica elettronica del paziente.
  Tale progetto è considerato cruciale per l'integrazione dell'intelligenza artificiale in medicina e anche necessario per la risoluzione di problemi di vario genere.

  L'ostacolo per l'adozione di questa nuova tecnologia medica è principalmente **il timore di una completa disumanizzazione della medicina** e almeno per il momento è destinata solo a risolvere problemi di tipo amministrativo e non approccia direttamente con il paziente e il suo stato d'animo[10; 15].

- **Fiducia**
  Una delle criticità correlate all'utilizzo dell'intelligenza artificiale nella pratica medica è il rischio che con il passare del tempo i medici possano **affidarsi eccessivamente** ai sistemi intelligenti.

  Questa fiducia potrebbe essere alimentata dall'idea che le nuove tecnologie siano considerate migliori della mente umana e una possibile conseguenza negativa a tale comportamento è il rischio di sviluppare una dipendenza da questi sistemi che con il passare del tempo potrebbe condurre alla riduzione del livello di competenza di medici, problema che risulterebbe più evidente e problematico se la tecnologia fallisse anche solo temporaneamente.

  Un esempio di questo fenomeno è un'analisi condotta della *City University of London* la quale ha evidenziato proprio che l'eccessivo affidamento a tecniche di machine learning ha influenzato le performance dei medici coinvolti [10].

- **Incertezza dei dati**
  Un altro aspetto critico causato da tecniche di machine learning è l'**incertezza che caratterizza molti fenomeni** in medicina.

  Questo aspetto è spesso trascurato nonostante sia ampiamente diffuso in medicina e potrebbe avere un impatto negativo sulla validità e attendibilita dei dati, sulle fasi di addestramento, sui test oppure sull'uso quaotidiano del sistema.

  L'aspetto fondamentale è che i dati necessari sono forniti dagli essere umani.
  Per esempio per addestrare un determinato algoritmo a riconoscere una certa patologia, come il melanoma attraverso immagini diagnostiche, deve ricevere in ingresso numerose fotografia, di casi già analizzati e diagnosticati da specialisti.

Questo aspetto genera un problema relativo alla validità dei dati, precisamente la discrepanza tra la qualità dei dati utilizzati per l'addestramento, che risulta essere sempre più elevata, e la qualità dei dati rilevati al momento della visita, la quale soddisfa raramente i requisiti richiesti.

Pertanto un sistema che riceve in input dati di scarsa qualità non è n grado di predirre un output corretto o del tutto affidabile [15].

- **Evoluzione**
  Un'ulteriore elemento critico riguarda il rischio di considrare non attendibile l'associazione tra un insieme di dati e una specifica diagnosi, tale rischio non può essere sottovalutato se si tratta di una diagnosi eseguita da un sistema intelligente in quanto la sua predizione si basa su dati considerati sicuramente veritieri.

  Il rischio peggiore, in questo caso, è la possibilità di condurre i medici a lavorare in modo autonomatico, senza l'utilizzo di queste nuove tecnologie, per evitare di commettere errori fatali, **impedendo in questo modo l'evoluzione dei sistemi intelligenti in un contesto medico** [15].

CAPITOLO 4

Metodologia

La **metodologia** è composta da una serie di metodi e tecniche che vengono applicate durante il processo di ricerca, nella *letteratura bianca(white literature)* e nella *letteratura grigia(gray literature)*, per ottenere un risultato valido.

4.1 Motore di ricerca

Per la *letteratura bianca(white literature)*, quindi per poter ricercare e selezionare i principali documenti scientifici pubblicati e il materiale accademico riguardante il problema di interesse, è stato utilizzato il motore di ricerca **Google Scholar** e i database di articoli accademici: **Scopus, IEEE Xplore, ACM digital library**.

Per la *letteratura grigia (gray literature)*, cioè i testi accademici diffusi dagli autori oppure da qualche ente privato o pubblico, è stato utilizzato **Google** come motore di ricerca.

4.2 Query di ricerca

Dopo aver stabilito il motore di ricerca per i due tipi di letteratura è stata identificata la **query di ricerca**, la quale rappresenta l'insieme di tutte le parole chiave che vengono utilizzate per ricercare fonti valide sia in letteratura bianca (white literatur) sia in letteratura la grigia (gray literature).
Dunque per tale indagine e per lo sviluppo dei capitoli è stata utilizzata specificatamente la seguente query di ricerca:
'FAIRNESS' AND ('MACHINE LEARNING' OR 'ML' OR 'AI' OR 'ARTIFICIAL INTELLIGENCE') AND ('IOT' OR 'INTERNET OF THINGS') AND 'MEDICINE'

4.3 Risultati ottenuti

Dopo aver utilizzato la query sopracitata all'interno del motore di ricerca **Google Scholar** sono stati generati **circa 30.000 risultati**, all'interno del database **Scopus** sono stati generati **circa 6.000 risultati**, all'interno del database **IEEE Xplore** sono stati generati **circa 15.000 risultati** e all'interno del database **ACM digital library** sono stati generati **circa 80.000 risultati**

4.4 Filtri applicati

Ai risultati ottenuti dall'utilizzo della query di ricerca è stato applicato un **filtro sulla data** per poter visualizzare facilmente i documenti con una data di pubblicazione più recente e a seguire sono stati scartati i documenti che dall'**abstract iniziale** non risultavano coerenti con l'argomento trattato.

4.5 Criteri di esclusione e inclusione

Successivamente è stato svolto un processo di **selezione** che consiste nell'eseguire una cernita dei risultati ottenuti utilizzando i criteri di *inclusione e di esclusione*.

Infatti tali criteri sono utilizzati per analizzare e stabilire se un determinato documento fosse abbastanza affidabile per essere considerato oppure se dovesse essere scartato.

Nella tabella 4.1 sono riportati tutti i criteri utilizzati per valutare ed eventualmente prendere in considerazione gli articoli riscontrati in letteratura *(white literatur)*.
La prima riga elenca tutti i criteri utilizzati per considerare un determinato docuento (i criteri di inclusione) mentre laseconda riga elenca tutti i criteri utilizzati per scartare un documento (i criteri di esclusione).

4.6 Risultati ottenuti

Dopo aver eseguito la selezione sono stati presi in considerazione **circa ventidue documenti**, sono stati denominati con **A#** e successivamente inseriti nella **prima bibliografia**.

Tabella 4.1: Criteri di inclusione e di esclusione

Criteri di inclusione	Nome dell'autore specificato
	Redatto in lingua inglese
	Coerente con gli argomenti trattati
	Fonte specificata
	Data di pubblicazione successiva al 2018
	Alta comprensibilità
	Alta credibilità del documento
	Argomento di interesse trattato minuziosamente
Criteri di esclusione	Nome dell'autore non specificato
	Non redatto in lingua inglese
	Non coerente con gli argomenti trattati
	Fonte non specificata
	Data di pubblicazione anteriore al 2018
	Scarsa comprensibilità
	Scarsa credibilità del documento
	Argomento di interesse trattato sommariamente

CAPITOLO 5

Analisi dei risultati

Il presente studio si è posto l'obiettivo di identificare i problemi di equità nei sistemi di machine learning e analizzare l'esistenza di tali problemi anche all'interno del settore sanitario.

In primo luogo, è stato identificato all'interno di ogni documento considerato il pensiero che gli autori avessero riguardante l'equità e il pregiudizio. Sono state proposte diverse definizioni di tale concetto, da un punto di vista filosofico e storico nelle aree dell'istruzione dell'apprendimento automatico, da un punto di vista statistico chiedendo la parità di alcune misure statistiche tra dei gruppi considerati e da un punto di vista individuale richiedendo vincoli che si legano a coppie specifiche di individui piuttosto che a un gruppo.

In secondo luogo, è stato analizzato il metodo di ricerca utilizzato. Tuttavia, solo all'interno di circa il 23% dei documenti è stato esplicitato tale metodo ed inoltre è risultato essere il medesimo. Infatti inizialmente è stata eseguita la raccolta dei dati con l'identificazione di eventuali dataset, successivamente è stata svolta una ricerca sistematica della letteratura tramite una query formata da parole chiave ed in conclusione una selezione finale del materiale.

Per quanto riguarda i dataset per i problemi di equità individuati durante la ricerca, sono stati esaminati da circa il 23% dei documenti selezionati solo il COMPAS Dataset utilizzato nei tribunali, da circa il 15% dei documenti il Ricci Dataset utilizzato per le promozioni lavorative, l'Adult Dataset utilizzato per il reddito annuo, il German Credit Dataset utilizzato per la valutazione del rischio, ed infine da circa l'8% dei documenti sono stati discussi il Bank Marketing Dataset utilizzato per il marketing bancario, Dutch Census Dataset utilizzato per classificare l'occupazione delle persone, KDD Census-Income Dataset utilizzato per il reddito totale, OULAD dataset utilizzato per prevedere il successo degli studenti.

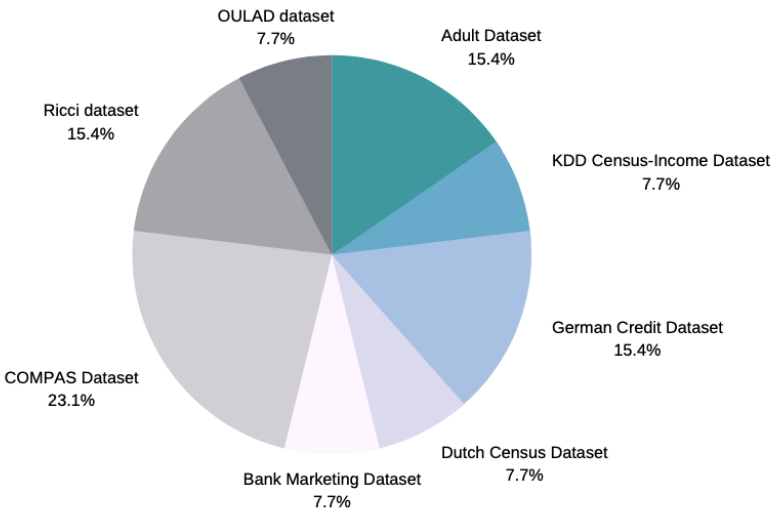

Figura 5.1: Dataset per problemi di equità

Per quanto riguarda tali documenti analizzati e studiati, per circa il 54%, affrontano temi di discriminazione e pregiudizi provocati da sistemi intelligenti in settori ordinari e per circa il 45% affrontano tali problemi esclusivamente nel settore medico.

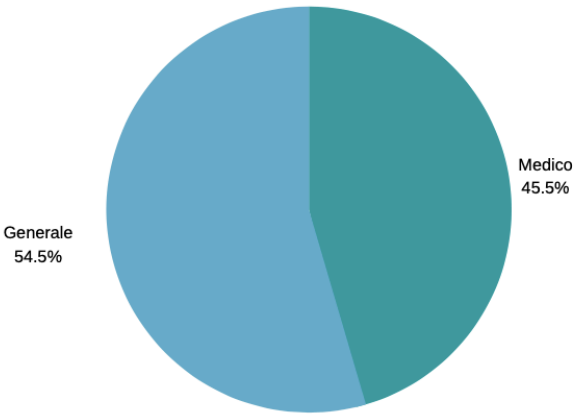

Figura 5.2: Tematiche affrontate dai documenti

Per il primo caso, sono stati analizzati principalmente, da circa il 15% delle fonti, i sistemi utilizzati per i consulenti intelligenti come Alexa o Siri e i sistemi di autenticazione più comuni, mentre da circa il 10% sono stati analizzati i moderni sistemi utilizzati al giorno d'oggi per la valutazione dei lavoratori e degli studenti, gli strumenti per eseguite il riconoscimento facciale e vocale.
Infine da circa il 5% dei documenti sono anche stati esaminati i sistemi per l'analisi delle immagini, i sistemi dei Robo-advisor, comuni consulenti finanziari intelligenti, i sistemi che governano i veicoli autonomi, i sistemi utilizzati nell'assistenza all'infanzia ed infine alcuni strumenti utilizzati nei tribunali per emettere il verdetto.

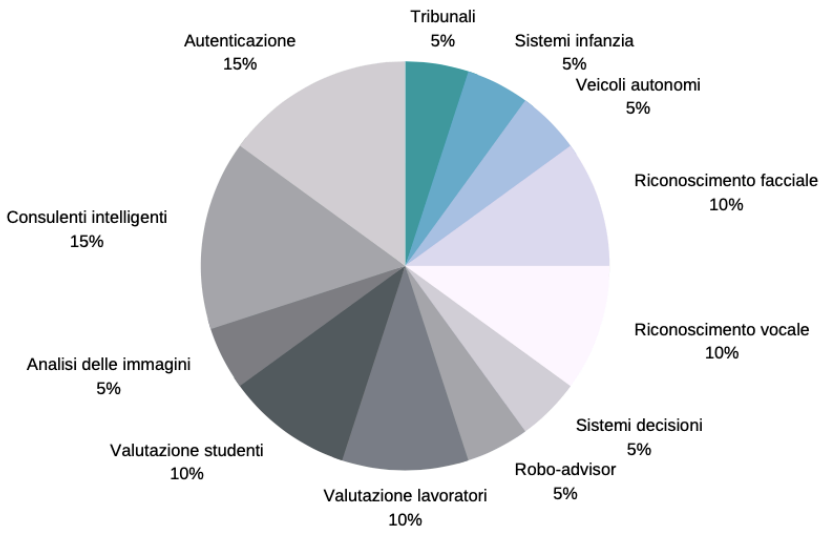

Figura 5.3: Discriminazioni dei sistemi intelligenti in settori ordinari

Per quanto riguarda il secondo caso, sono stati studiati, da circa il 35% delle fonti, i sistemi intelligenti utilizzati per assistere i medici durante la diagnosi delle malattie, da circa il 20%, la creazione e l'utilizzo di dispositivi indossabili ed i sistemi di monitoraggio intensivo del paziente. Da circa il 10% sono state analizzate le cartelle cliniche elettroniche ed infine da circa il 5% delle fonti sono state analizzate le discriminazioni e i pregiudizi nei sistemi utilizzati per la comunicazione con il paziente, per l'analisi del tasso di recidiva e per l'adozione della chirurgia robotica.

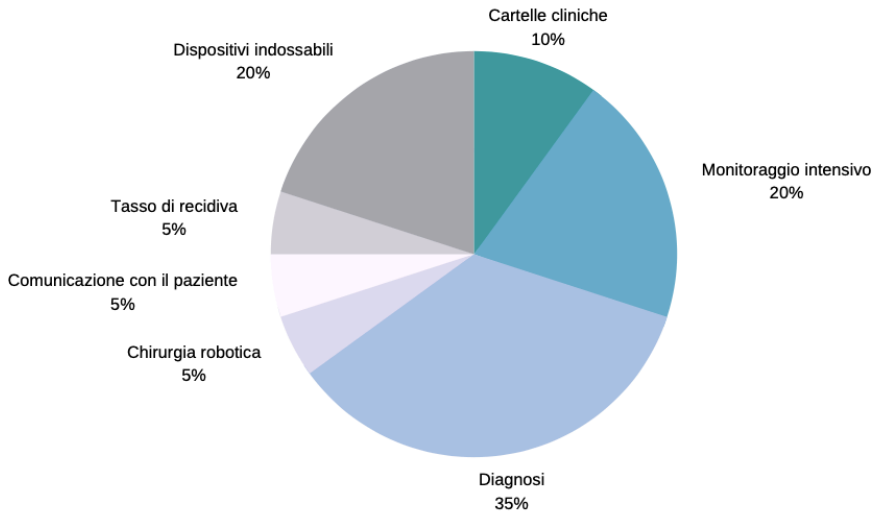

Figura 5.4: Discriminazioni dei sistemi intelligenti nel settore medico

Inoltre, sono stati analizzati in modo dettagliato i sistemi intelligenti e i relativi pregiudizi in svariati rami della medicina: nella cardiologia da circa il 30% degli autori, nella neurologia, nella gastroenterologia, nell'endocrinologia e nella medicina polmonare da circa il 12%, nella nefrologia da circa il 18% e nella Fisioterapia da circa il 6%.

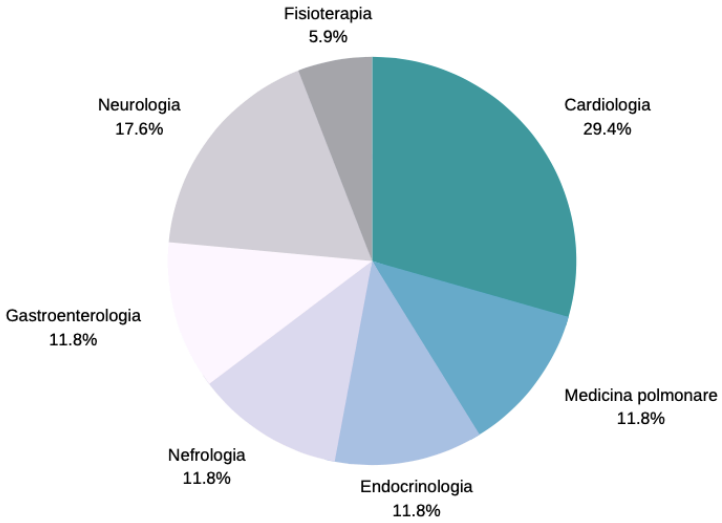

Figura 5.5: Discriminazioni dei sistemi intelligenti in specifici rami della medicina

Analizzato quanto detto sopra, dalla ricerca è emerso che tra tutti gli attributi considerati, quelli nei quali si evidenzia una maggiore discriminazione nei modelli di apprendimento automatico sono principalmente il sesso e l'etnia, a seguire lo status sociale, la religione e la modalità di comportamento del singolo individuo.

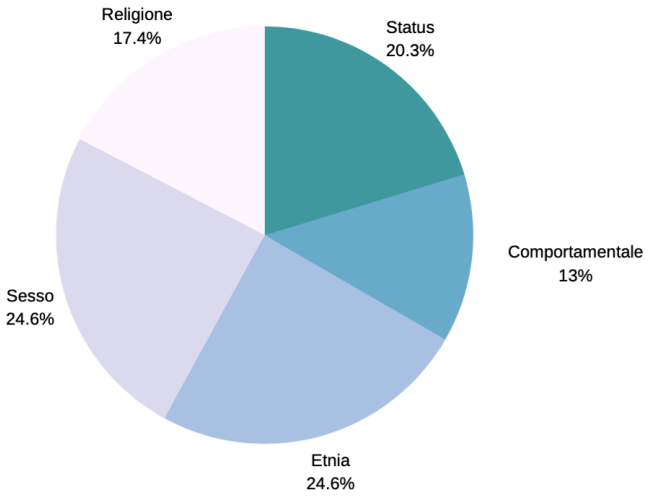

Figura 5.6: Attributi discriminati

Inoltre sono stati individuati dei vantaggi e degli svantaggi provocati dall'applicazione di questi algoritmi intelligenti nei settori comuni.

Tuttavia, le caratteristiche riscontrate durante la ricerca mettono in luce principalmente gli aspetti negativi derivati dall'utilizzo di tali tecnologie, infatti circa il 23% delle ricerche svolte afferma che il problema principale è la perdita di privacy provocata molto spesso dall'uso costante di sistemi intelligenti.
Circa il 15% afferma che gli svantaggi principali siano l'iniquità sociale o culturale che generalmente possono provocare tali sistemi, la possibilità di generare un qualsiasi tipo di malfunzionamento in qualunque momento e la possibilità di causare una possibile perdita di lavoro per il singolo individuo.
Un gruppo ridotto, circa l'8% sostiene che gli svantaggi preoccupanti sono l'eccessiva speranza riposta in tali dispositivi da parte delle persone comuni, la possibile influenza negativa sullo stato socioeconomico, la sottovalutazione dell'eccessiva complessità di utilizzo di tali dispositivi oppure, nel settore sanitario, la perdita totale del rapporto tra medico e paziente.

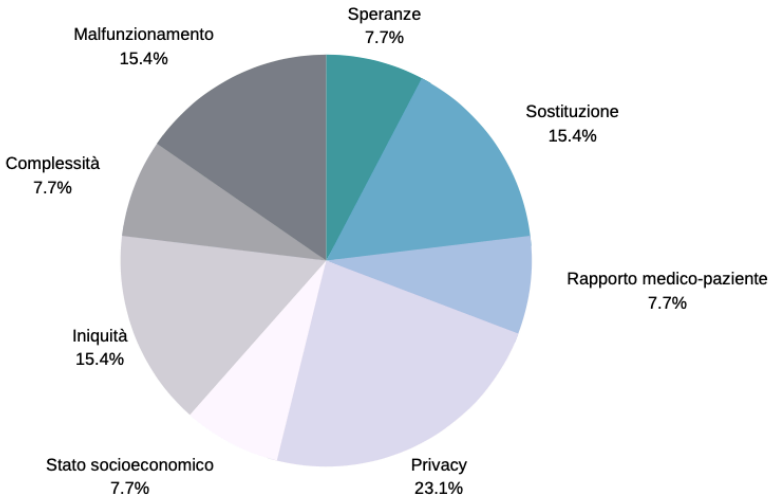

Figura 5.7: Svantaggi degli algoritmi intelligenti

Per quanto riguarda le caratteristiche positive evidenziate, circa il 33% delle ricerche svolte afferma che il vantaggio principale offerto da queste nuove tecnologie è stato il miglioramento della qualità di vita del singolo individuo quotidianamente. Circa il 17% sostiene che i vantaggi sono sia quotidiani, come la possibilità di analizzare meticolosamente alcuni dati oppure l'ottimizzazione di singoli processi comuni, sia sanitaria con un'assistenza costante da parte dei dispositivi intelligenti offerta ai medici durante azioni più complesse come quelle chirurgiche ma anche azioni abituali come una diagnosi più accurata.

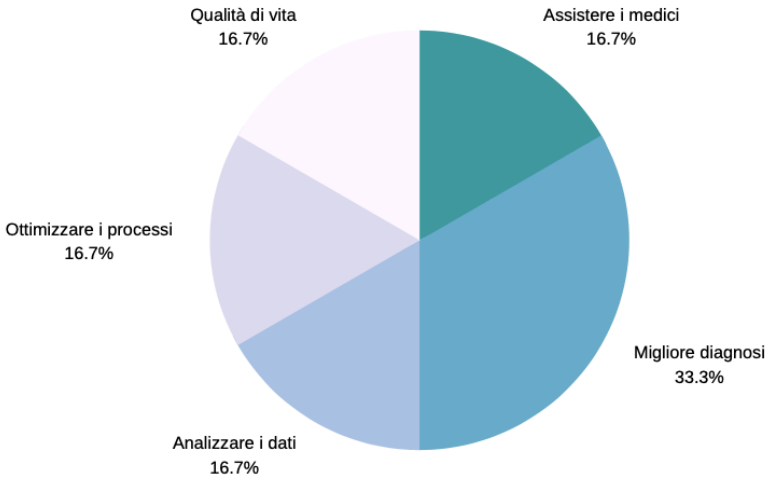

Figura 5.8: Vantaggi degli algoritmi intelligenti

Un aspetto interessante ma discusso solo nel 15% circa dei documenti è stata l'incapacità di proporre soluzioni algoritmiche per il quadro sociotecnico, questo aspetto è stato scaturito dalla mancata comprensione di concetti sociali fondamentali e di come l'inserimento della tecnologia in un sistema sociale esistente possa influenzare e modificare, sia negativamente sia positivamente, i comportamenti e i valori della popolazione.
Inoltre, nella ricerca è stata rimarcata l'idea che gli algoritmi non tengono conto delle complesse relazioni tra fattori biologici, ambientali e sociali provocando in questo modo differenze sostanziali nella popolazione e notevoli iniquità.

Oltre a ciò, sono state anche identificate delle metriche e dei criteri per valutare la presenza di equità all'interno degli algoritmi intelligenti. I criteri di equità astratta, le metriche di equità di gruppo, le metriche basate sulla parità, le metriche basate su matrice di confusione, le metriche basate sulla calibrazione, le metriche basate sul punteggio e le metriche di equità individuale.

Nel corso della ricerca sono stati identificati diversi processi per mitigare le ingiustizie e alcuni metodi alternativi per risolvere la disparità provocate dagli algoritmi intelligenti.

Tra i processi per mitigare le ingiustizie vi sono: la classificazione binaria che è l'approccio più comune, il Blinding l'approccio per rendere un classificatore immune a una o più variabili sensibili, i metodi causali i quali riconoscono che i dati su cui vengono addestrati i modelli spesso riflettono una qualche forma di discriminazione, i metodi di campionamento che creano campioni per il training di algoritmi robusti cercando di correggere i dati, l'identificazione dei gruppi di dati che sono significativamente svantaggiati, la trasformazione che consiste in una nuova rappresentazione dei dati, la ri-etichettatura e perturbazione che modificano la variabile dipendente o modificano la distribuzione di uno o più variabili direttamente nei dati di addestramento, la ri-pesatura che assegna pesi alle istanze dei dati di addestramento lasciando i dati invariati, la regolarizzazione ed eventualmente l'ottimizzazione dei vincoli che aggiungono uno o più termini sanzionatori che penalizzano il classificatore per pratiche discriminatorie, l'apprendimento di tipo contraddittorio in un processo di rilevamento di campioni di dati falsificati, la calibrazione in un processo per garantire che la proporzione di previsioni positive sia uguale alla proporzione di esempi positivi, la soglia un processo basato sul fatto che le decisioni discriminatorie sono spesso prese vicino ai confini del processo decisionale a causa del pregiudizio e che gli esseri umani applicano regole di soglia quando prendono decisioni.

Invece per quanto riguarda i metodi alternativi per risolvere la disparità dell'apprendimento automatico circa il 28% delle ricerche svolte affermano che per ottenere la risoluzione del problema si debba avere una parità demografica, una parità del tasso predittivo, una parità individuala, una parità di prestazioni dei modelli, una parità dei gruppi considerati ed avere un contesto di ricerca e dei gruppi di confronto equi.

Circa il 20% dichiara che la base per ottenere l'equità sia l'implementazione di un programma progettato per non considerare l'etnia oppure progettare l'algoritmo specifico in modo da far incorporare esplicitamente o implicitamente l'etnia quindi ottenere comunque algoritmi neutri.

Circa il 16% sostiene che bisognerebbe concentrarsi sul modello linguistico, incorporare le parole, potenziare la traduzione automatica o eseguire una buona formulazione del problema.

Circa il 16% attesta che la chiave per eliminare la disparità sia nei dati; infatti, propone un ri-campionamento di questi, l'utilizzo dei dati imparziali oppure eseguire un pretrattamento e un'analisi dei dati più accurata.

Infine, circa il 4% ritiene che la soluzione sia tenere conto del ciclo di Feedback, considerare gli aspetti del processo di ingegneria del software, avere una trasparenza per lo sviluppo del modello, mappare i requisiti per l'equità dei diversi problemi oppure utilizzare delle metriche per valutare e soddisfare l'equità.

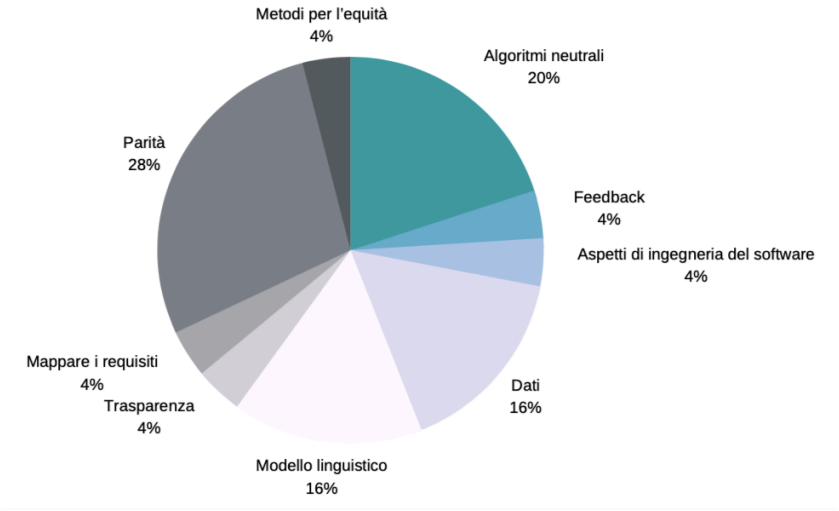

Figura 5.9: Metodi alternativi per risolvere la disparità

Infine, sono stati identificati anche degli algoritmi utilizzati per cercare di risolvere i problemi di equità generati in sistemi intelligenti. Tuttavia solo nel 15% dei documenti sono stati esplicitati tali algoritmi, il primo individuato è SMOTE utilizzato per sintetizzare in modo intelligente e generare nuovi dati per far crescere la classe minoritaria, successivamente sono stati descritti gli algoritmi Calders e Verwer, Feldman, Kamishima, Zafar ed infine l'algoritmo Unchoke utilizzato nel Bootstrapping nel quale ogni nodo che cerca di entrare a far parte della rete dovrebbe dimostrare rapidamente che è genuino e potenzialmente onesto.

CAPITOLO 6

Conclusioni e sviluppi futuri

Con le ricerche svolte è stato raggiunto l'obiettivo principale di evidenziare i problemi di equità che gli algoritmi utilizzati nelle tecnologie intelligenti possono generare, cioè la loro capacità di denigrare o di emarginare una parte di popolazione per lo stato sociale, la religione, il sesso o l'etnia.

Quanto ultimato durante tutto questo studio può essere riepilogato nei seguenti punti, i quali rappresentano i temi cruciali analizzati:

- Stabilire il significato di equità in letteratura ed il concetto di giudizio negli esseri umani e nelle macchine.
- Individuare le metriche più comuni in letteratura utilizzate per quantificare l'equità e le tecniche di valutazione maggiormente considerate.
- Esaminare le motivazioni che stimolano l'interesse per l'equità.
- Studiare i dataset più adatti a questo tipo di problemi.
- Identificare i dispositivi che generano tale iniquità in ambienti comuni e in ambienti sanitari.
- Esaminare altre problematiche provocate dai sistemi intelligenti solo nel settore sanitario.
- Ricercare degli esempi significativi che evidenziassero tale pregiudizio nella società moderna.

Inoltre con lo scopo di rendere tale studio completo ed essere quanto più meticolosi e accurati possibili nella ricerca e nell'analisi di informazioni attendibili in letteratura, sono state eseguite più ricerche su differenti motori di ricerca. Infatti, dopo aver formato la query di ricerca da utilizzare, questa è stata inizialmente applicata al motore di ricerca Google Scholar e in un secondo momento ai database di articoli accademici come Scopus, IEEE Xplore e ACM digital library. In questo modo è stato possibile confrontare i dati raccolti da diverse fonti e appurare la presenza di risultati molto simili tra di loro.

Da questo è stato possibile concludere che:

- Le metriche più comuni sono di equità di gruppo, basate sulla parità, basate su matrice di confusione e di equità individuale.
- Le motivazioni che stimolano l'interesse del singolo individuo per l'equità sono strumentali, relazionali e morali.
- Gli algoritmi intelligenti sono maggiormente propensi a discriminare secondo il sesso e l'etnia del singolo individuo.
- La causa principale della discriminazione è la qualità e quantità dei dati di addestramento dell'algoritmo.
- Altre principali problematiche esterne dovute all'utilizzo di queste tecnologie sono la perdita di privacy e l'iniquità culturale.
- Il vantaggio principale è il miglioramento costante della qualità di vita del singolo individuo.

Tuttavia, trattandosi di sistemi in continua evoluzione queste caratteristiche riscontrate potrebbero migliorare o peggiorare con il tempo ed oltretutto, considerando che i problemi di equità sono diventati oggetto di studio da poco e che le informazioni a disposizione sono ancora limitate, sicuramente si verificherà una modifica o un avanzamento dei dati raccolti.

6.1 Sviluppi futuri

- **Ricerca:** Per migliorare quanto descritto è possibile eseguire quotidianamente una ricerca approfondita sugli sviluppi legati all'introduzione di tecnologie intelligenti all'interno dei sistemi utilizzati abitualmente. Al fine di avanzare giornalmente le conoscenze sul tema e poter ricercare, studiare e proporre eventuali soluzioni valide.
- **Atteggiamento:** È necessario approfondire gli atteggiamenti, positivi e negativi, che alcune persone adottano nei confronti della tecnologia e i limiti che essi impongono all'avanzamento della tecnologia. Infatti, si dovrebbe verificare se, con la completa digitalizzazione dei sistemi, esse siano disposte ad accettare le relative problematiche causate per godere dei vantaggi quotidiani offerti oppure se preferiscano adoperare sistemi obsoleti per garantire la propria sicurezza.
- **Cultura:** Per dare una visione competa e accurata dei problemi causati dai sistemi intelligenti è necessario studiarli differenziando le varie culture a cui tali sistemi sono applicati. Infatti, sarebbe interessante accertare come il concetto di equità e pregiudizio varia a seconda della comunità considerata. Tale pensiero influenzerebbe tutta la ricerca svolta sui sistemi intelligenti.
- **Medicina:** Verificare gli avanzamenti quotidiani della tecnologia medica, i benefici che essi comportano per determinati pazienti e le possibili soluzioni proposte per risolvere eventuali problemi. Nello specifico si potrebbe approfondire l'utilizzo di alcune tecnologie avanzate su pazienti con determinate malattie terminali per verificare quanto e come la nuova tecnologia e l'intelligenza artificiale possano intervenire nel ciclo di vita umano considerato, per il momento, limitato ma con tali sistemi potrebbe magari divenire interminabile.

Ringraziamenti

Desidero ringraziare il professore Fabio Palomba, relatore di questa tesi di laurea, in primo luogo per la capacità di stimolare, attraverso le sue lezioni e i suoi insegnanti, il mio interesse per l'argomento trattato, in secondo luogo per l'aiuto fornitomi da lui e dal Dott. Giammaria Giordano, correlatore di questa tesi, per la disponibilità, la precisione e la dedizione fornitami durante tutta la stesura di questo elaborato. Senza la vostra guida costante e il vostro aiuto non sarei riuscita a raggiungere questo livello di conoscenza e questo traguardo fondamentale per la mia carriera.

Un ringraziamento speciale va ai miei genitori.

Ringrazio mio padre Pietro, il quale è stato il mio punto di riferimento. Attraverso il suo percorso di vita mi ha dimostrato e insegnato che per raggiungere gli obiettivi e superare le difficoltà bisogna lavorare duramente e con costanza.

Ringrazio mia madre Luisa, la quale mi ha sostenuta durante tutta la mia vita e in tutte le mie scelte. Ha creduto in me prima ancora che lo facessi io ma soprattutto mi ha insegnato che non esistono sconfitte che non possano trasformarsi in successi.

Voi siete il mio esempio di vita da seguire. Grazie ai vostri sacrifici e al vostro sostegno sono riuscita a raggiungere questo traguardo.

Ringrazio due persone meravigliose che ho avuto la fortuna di avere al mio fianco, le mie sorelle, coloro che attraverso la conoscenza e la passione per la medicina mi hanno trasmesso l'ispirazione adatta per redigere una parte significativa di questa tesi.

Ringrazio mia sorella Annalina, la quale, nonostante la distanza, mi ha ascoltata e sostenuta in ogni occasione. Attraverso la sua determinazione ho capito che si può ottenere tutto se lo si desidera ardentemente.

Ringrazio mia sorella Isabella, la quale mi supporta giorno dopo giorno nelle decisioni e nelle difficoltà della vita. Mi ha concesso un sorriso ad ogni sconfitta ed una spalla costante su cui appoggiarmi.

Siete da sempre un esempio da ammirare ed imitare. Grazie per essere sempre al mio fianco e dalla mia parte, senza i vostri insegnamenti, giudizi e consigli non sarei riuscita a diventare la persona che sono oggi.

Infine, vorrei ringraziare gli amici con cui sono cresciuta Christian, Elena e Serena ma anche i colleghi che ho avuto il piacere di conoscere durante questo percorso universitario Damiana, Pio e Alessandro, ognuno di voi a modo suo ha avuto un peso importante nella mia crescita e nella mia formazione. E se sono riuscita ad arrivare fin qui e anche grazie alle esperienze vissute e al sostegno che mi avete dato ogni giorno.

Grazie a tutti voi.

Bibliografia

[A1] Muhammad Aurangzeb Ahmad, Carly Eckert, Christine Allen, Vikas Kumar, Juhua Hu, and Ankur Teredesai. Fairness in healthcare ai. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 554–555. IEEE, 2021.

[A2] Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3529–3530, 2020.

[A3] Giovanni Briganti and Olivier Le Moine. Artificial intelligence in medicine: today and tomorrow. *Frontiers in medicine*, 7:27, 2020.

[A4] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

[A5] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[A6] Richard Ribón Fletcher, Audace Nakeshimana, and Olusubomi Olubeko. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Frontiers in Artificial Intelligence*, 3:116, 2021.

[A7] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

[A8] Kevin B Johnson, Wei-Qi Wei, Dilhan Weeraratne, Mark E Frisse, Karl Misulis, Kyu Rhee, Juan Zhao, and Jane L Snowdon. Precision medicine, ai, and the future of personalized health care. *Clinical and translational science*, 14(1):86–93, 2021.

[A9] Fabrice Jotterand, Alexandre Erler, and Vincent C Müller. Ai as ia: The use and abuse of artificial intelligence (ai) for human enhancement through intellectual augmentation (ia). 2021.

[A10] Chenglu Li, Wanli Xing, and Walter L Leite. Do gender and race matter? supporting help-seeking with fair peer recommenders in an online algebra learning platform. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 432–437, 2022.

[A11] Melissa D McCradden, Shalmali Joshi, Mjaye Mazwi, and James A Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, 2020.

[A12] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[A13] D Douglas Miller and Eric W Brown. Artificial intelligence in medical practice: the question to the answer. *The American journal of medicine*, 131(2):129–133, 2018.

[A14] Francis Nwebonyi, Rolando Martins, and Manuel Eduardo Correia. Security and fairness in iot based e-health system: A case study of mobile edge-clouds. pages 318–323, 10 2019.

[A15] Annunziata Paviglianiti and Eros Pasero. Vital-ecg: a de-bias algorithm embedded in a gender-immune device. In *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*, pages 314–318. IEEE, 2020.

[A16] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *arXiv preprint arXiv:2110.00530*, 2021.

[A17] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[A18] Amirhossein Rasooli, Hamed Zandi, and Christopher DeLuca. Re-conceptualizing classroom assessment fairness: A systematic meta-ethnography of assessment literature and beyond. *Studies in Educational Evaluation*, 56:164–181, 2018.

[A19] Georg Starke, Eva De Clercq, and Bernice S Elger. Towards a pragmatist dealing with algorithmic bias in medical machine learning. *Medicine, Health Care and Philosophy*, 24(3):341–349, 2021.

[A20] IA Sulaimon, A Ghoneim, and M Alrashoud. Adaptation of machine learning based fairness algorithm for real time decision in autonomous systems.

[A21] Andreas Tsamados, Nikita Aggarwal, Josh Cowls, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi. The ethics of algorithms: key problems and solutions. *AI & SOCIETY*, pages 1–16, 2021.

[A22] Zeeshan Zulkifl, Fawad Khan, Shahzaib Tahir, Mehreen Afzal, Waseem Iqbal, Abdul Rehman, Saqib Saeed, and Abdullah M Almuhaideb. Fbashi: Fuzzy and blockchain-based adaptive security for healthcare iots. *IEEE Access*, 2022.

Bibliografia

[1] Come funzionano gli algoritmi di machine learning. 2021.

[2] Il tribunale condanna deliveroo: l'algoritmo che valuta i rider è discriminatorio. 2021.

[3] Intelligenza artificiale in campo medico: quali sono le applicazioni oggi. 2021.

[4] L'intelligenza artificiale aiuta la riabilitazione motoria. 2021.

[5] Se l'algoritmo è discriminatorio. 2021.

[6] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*, 2021.

[7] Laurie J Barclay, Michael R Bashshur, and Marion Fortin. Motivated cognition and fairness: Insights, integration, and creating a path forward. *Journal of Applied Psychology*, 102(6):867, 2017.

[8] Elisa Bertino. Data security and privacy in the iot. In *EDBT*, volume 2016, pages 1–3, 2016.

[9] Nicoletta Boldrini. Algoritmi e discriminazione di genere. facebook ci ricasca. 2021.

[10] Giovanni Briganti and Olivier Le Moine. Artificial intelligence in medicine: today and tomorrow. *Frontiers in medicine*, 7:27, 2020.

[11] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

[12] J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women.

[13] Wei Du and Xintao Wu. Robust fairness-aware learning under sample selection bias. *arXiv preprint arXiv:2105.11570*, 2021.

[14] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. *arXiv preprint arXiv:2110.00857*, 2021.

[15] RAFFAELE RASOINI GIAN FRANCO GENSINI FEDERICO CABITZA, CAMILLA ALDERIGHI. Potenziali conseguenze inattese dell'uso di sistemi di intelligenza artificiale oracolari in medicina. 2017.

[16] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE, 2020.

[17] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*, pages 498–510, 2017.

[18] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

[19] Yair Horesh, Noa Haas, Elhanan Mishraky, Yehezkel S Resheff, and Shir Meir Lador. Paired-consistency: an example-based model-agnostic approach to fairness regularization in machine learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 590–604. Springer, 2019.

[20] Alexey Ignatiev, Martin C Cooper, Mohamed Siala, Emmanuel Hebrard, and Joao Marques-Silva. Towards formal fairness in machine learning. In *International Conference on Principles and Practice of Constraint Programming*, pages 846–867. Springer, 2020.

[21] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Model-based and actual independence for fairness-aware classification. *Data mining and knowledge discovery*, 32(1):258–286, 2018.

[22] Dana Pessach and Erez Shmueli. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*, 2020.

[23] Silvia Peviani. Intelligenza artificiale e robotica nella fisioterapia. 2017.

[24] Tai Le Quy, Arjun Roy, Gunnar Friege, and Eirini Ntoutsi. Fair-capacitated clustering. *arXiv preprint arXiv:2104.12116*, 2021.

[25] AN Ramesh, Chandra Kambhampati, John RT Monson, and PJ Drew. Artificial intelligence in medicine. *Annals of the Royal College of Surgeons of England*, 86(5):334, 2004.

[26] Vincenzo Randazzo, Eros Pasero, and Silvio Navaretti. Vital-ecg: A portable wearable hospital. In *2018 IEEE Sensors Applications Symposium (SAS)*, pages 1–6. IEEE, 2018.

[27] Shirin Riazy and Katharina Simbeck. Predictive algorithms in learning analytics and their fairness. *DELFI 2019*, 2019.

[28] Goce Ristanoski, Wei Liu, and James Bailey. Discrimination aware classification for imbalanced datasets. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1529–1532, 2013.

[29] Alessandro Rizzuto. Medicina e intelligenza artificiale: come le macchine possono migliorare la nostra salute. 2020.

[30] Michael Siegrist, Melanie Connor, and Carmen Keller. Trust, confidence, procedural fairness, outcome fairness, moral conviction, and the acceptance of gm field experiments. *Risk Analysis: An International Journal*, 32(8):1394–1403, 2012.

[31] Domedica s.r.l. Intelligenza artificiale e medicina, tra futurismo e concrete realtà. 2017.

[32] IA Sulaimon, A Ghoneim, and M Alrashoud. Adaptation of machine learning based fairness algorithm for real time decision in autonomous systems.

[33] Ana Valdivia, Javier Sánchez-Monedero, and Jorge Casillas. How fair can we go in machine learning? assessing the boundaries of accuracy and fairness. *International Journal of Intelligent Systems*, 36(4):1619–1643, 2021.

[34] Enrico Verga. Intelligenza artificiale, la discriminazione da parte degli algoritmi è un pericolo reale. 2021.

[35] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[36] Kent Walker. Ai for social good in asia pacific. *Around the Globe*, 2018.

[37] Raffaele Zenti. Più odio, più profitti: la dura lotta fra intelligenza artificiale e stupidità naturale. 2021.

[38] Jie M Zhang and Mark Harman. "ignorance and prejudice" in software fairness. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 1436–1447. IEEE, 2021.