EBOOKECM JOURNAL
N. 5 - GEN 2023

IA IN MEDICINA

Applicazioni e Prospettive

ARTICOLI SELEZIONATI E TRADOTTI DA

Medical Image Analysis // British Journal of Cancer // Journal of Medical Internet Research // Artificial Intelligence in Medicine // International Journal of Environmental Research and Public Health // Biosensors // Journal of Personalized Medicine // Diagnostics // Life // Frontiers in Artificial Intelligence

La presente pubblicazione è accreditata come **corso ECM FAD** solo attraverso apposita registrazione su www.ebookecm.it

EBOOKECM JOURNAL

Titolo: Intelligenza Artificiale in Medicina. Applicazioni e prospettive
Curatela: Marco Cascella
Traduzioni: Giuditta Spassini
Editing e copertina: Attilio Scullari
Concept copertina: Licia Casula
Direzione editoriale: Alessandra Pontis, Mario Marcello Verona
Supervisione scientifica: Carlo Duò, Alessandra Pontis
Data Pubblicazione: Gennaio 2023

COLLANA EBOOKECM
EBOOK PER L'EDUCAZIONE CONTINUA IN MEDICINA © 2022

ISBN: 9788831253932
ISSN: 2785-2911

BOOKIA SRL. Servizi di editoria accreditata, Piazza Deffenu 12, 09125 Cagliari.

INDICE

INTRODUZIONE

"Possono le macchine pensare?" La frase di Alan Turing è riportata nel suo articolo, "Computing Machinery and Intelligence" pubblicato sulla rivista Mind nel 1950. Il celeberrimo scritto in cui il matematico inglese descriveva l'esperimento che, inizialmente indicato come gioco dell'imitazione, prenderà il suo nome (esperimento di Turing) ha, di fatto, segnato l'inizio dell'era dell'Intelligenza Artificiale. In verità, il termine Intelligenza Artificiale è stato coniato dall'informatico John McCarthy che, nel 1956, organizzò la famosa conferenza di Dartmouth (College di Hannover) nel New Hampshire in America. Nel corso della conferenza, denominata Summer Research Project on Artificial Intelligence, lo stesso John McCarthy e altri nove ricercatori affrontarono i temi delle reti neurali, e lo sviluppo di algoritmi e sistemi di calcolo finalizzati alla conoscenza (e imitazione) dei processi propri dell'intelligenza "naturale": la nuova disciplina dell'Intelligenza Artificiale era stata programmaticamente fondata. Nell'introdurre la conferenza, McCarthy affermò che: "Lo studio deve procedere sulla base della congettura che ogni aspetto dell'apprendimento o qualsiasi altra caratteristica dell'intelligenza possa, in linea di principio, essere descritto in modo così preciso che sia possibile costruire una macchina per simularlo (Intelligenza Artificiale, n.d.r.)".

Contrariamente alla falsa convinzione che identifica l'Intelligenza Artificiale con tecnologie futuristiche e robot senzienti capaci di comprendere e decidere le azioni da compiere e di convivere con gli esseri umani ed eventualmente soppiantarli, l'Intelligenza Artificiale è più propriamente definita come la branca dell'informatica che, attraverso la progettazione e programmazione di sistemi hardware e software, consente di dotare le macchine di determinate caratteristiche che sono tipicamente

considerate umane. Tali proprietà includono, ad esempio, percezioni visive, spazio-temporali e processi decisionali di supporto. In questo ambito nosografico, la vera sfida della Intelligenza Artificiale è di riprodurre un percorso funzionale che è il cardine dell'intelligenza naturale ed è dinamicamente strutturato intersecando i contenuti di una conoscenza non sterile, la capacità di prendere decisioni e risolvere problemi non solo secondo la logica, ma adattandosi ai contesti operativi. Tuttavia, le due forme di intelligenza, naturale (biologica o al carbonio) e artificiale (al silicio), possono integrarsi, ma non essere interscambiabili. È doveroso sottolineare, infatti, le principali differenze in merito a velocità operativa e domini di azione. La velocità operativa è enormemente superiore nei sistemi artificiali, poiché mentre in questi ultimi i segnali si propagano quasi alla velocità della luce, negli esseri umani la velocità di conduzione degli impulsi nervosi è molto più lenta. Questa caratteristica comporta che la connettività e i fenomeni comunicativi sono particolarmente sviluppati nelle macchine rispetto agli umani. D'altro canto, in confronto alle macchine, i cervelli umani ottimizzano il dispendio energetico e, soprattutto, sono particolarmente efficienti nei compiti percettivi motori: il processo evolutivo degli esseri umani ha preferito sviluppare le abilità finalizzate alla integrazione con l'ambiente, la riproduzione e la sopravvivenza, piuttosto che implementare la potenza del calcolo, l'archiviazione e la scalare elaborazione dei dati (paradosso di Moravec).

All'entusiasmo iniziale per questa affascinante nuova branca delle scienze è seguito il cosiddetto "inverno dell'Intelligenza Artificiale", locuzione utilizzata per indicare un periodo tra gli anni '70 e gli anni '90, in cui le limitate capacità della strumentazione disponibile (potenza di calcolo e ridotta acquisizione ed elaborazione dei dati) hanno condizionato lo sviluppo della disciplina. Successivamente, a partire dagli anni 2010 e grazie al progresso tecnologico, l'Intelligenza Artificiale ha avuto una incredibile fase di crescita. In questa "primavera dell'Intelligenza Artificiale" anche nel campo medico sono stati ottenuti importanti risultati. La digitalizzazione dei dati sanitari ha consentito, infatti, di

strutturare e analizzare big data e metadata che rappresentano una solida base per il funzionamento di algoritmi e reti neurali. È indubbio, oramai, che l'uso di modelli predittivi rappresenti un'importante opportunità in medicina e i vantaggi dell'Intelligenza Artificiale e delle sue branche come il Machine Learning si sostanziano in un significativo miglioramento dell'assistenza clinica ai pazienti, ma riguardano anche i processi organizzativi e i sistemi sanitari. Nella pianificazione dei percorsi di cura, per esempio, l'Intelligenza Artificiale risulta essere una risorsa quantomai utile per migliorare i flussi di lavoro ospedalieri, individuando le attività che richiedono priorità e fornendo un servizio adeguato alle esigenze del paziente. A titolo esemplificativo, l'applicazione più recente dell'Intelligenza Artificiale in ambito sanitario riguarda la pandemia di SARS-CoV-2, dove diverse strategie computazionali predittive sono state utilizzate al fine di migliorare la diagnosi, prevedere l'andamento dell'epidemia, identificare i pazienti a rischio di sviluppare forme severe di malattia e definire programmi terapeutici.

Marco Cascella

1 - L'INTELLIGENZA UMANA CONTRO QUELLA ARTIFICIALE

Tratto e tradotto da

E. (Hans). Korteling, G. C. van de Boer-Visschedijk, R. A. M. Blankendaal, R. C. Boonekamp e A. R. Eikelboom, *Human- versus Artificial Intelligence*, *TNO Human Factors*, Front. Artif. Intell., 25 March 2021, Sec. AI for Human Learning and Behavior.

https://doi.org/10.3389/frai.2021.622364

Le parti omesse dal curatore rispetto all'originale sono indicate dal segno **[...]**

L'intelligenza artificiale è uno degli argomenti più dibattuti al giorno d'oggi e sembra che ci sia in generale poca comprensione riguardo alle differenze e alle somiglianze tra intelligenza umana e intelligenza artificiale. Le discussioni su molti argomenti inerenti, come l'affidabilità, la spiegabilità e l'etica, sono frutto di concezioni antropocentriche e antropomorfe e, ad esempio, dalla ricerca di un'intelligenza simile a quella umana come ideale per l'intelligenza artificiale. Al fine di cercare un maggiore consenso e di consolidare possibili obiettivi di ricerca futuri, il presente documento indaga tre concetti sul tema delle somiglianze e differenze tra intelligenza umana e artificiale: 1) i vincoli principali dell'intelligenza umana (e artificiale), 2) l'intelligenza umana vista come una delle forme possibili di intelligenza generale e 3) l'elevato impatto potenziale di forme multiple (integrate) di applicazioni di IA ibride e ristrette. Per il momento, i sistemi di IA avranno qualità e capacità cognitive molto diverse da quelle dei sistemi biologici. Per questo motivo, una delle questioni più importanti è come possiamo utilizzare (e "collaborare" con) questi sistemi nel modo più efficace possibile. Per quali compiti e in quali condizioni è sicuro delegare le decisioni all'IA e quando è

necessario il giudizio umano? Come possiamo sfruttare i punti di forza specifici dell'intelligenza umana e artificiale? Come impiegare efficacemente i sistemi di IA per integrare e compensare i limiti intrinseci della cognizione umana (e viceversa)? Dobbiamo perseguire lo sviluppo di IA che collaborano con l'intelligenza umana (o simile all'umana) o concentrarci più sull'integrazione dei limiti umani? Per rispondere a queste domande gli esseri umani che lavorano con l'IA, o che ne definiscono le regole, devono sviluppare una comprensione adeguata dei meccanismi "psicologici" sottostanti all'IA. Quindi, per ottenere sistemi umani-IA ben funzionanti, dovrebbe essere affrontata meglio la *consapevolezza dell'intelligenza* negli esseri umani. A questo scopo viene proposto un primo quadro di riferimento a scopo educativo.

1. INTRODUZIONE: INTELLIGENZA ARTIFICIALE E UMANA, DUE MONDI DIVERSI

1.1 INTELLIGENZA ARTIFICIALE GENERALE AL LIVELLO UMANO

I recenti progressi dell'informatica e dell'IA possono consentire una maggiore coordinazione e integrazione tra uomo e tecnologia. Per questo motivo, è stata dedicata una certa attenzione allo sviluppo dell'IA *consapevole dell'uomo*, che si adatti come "membro di una squadra" alle potenzialità cognitive e ai limiti della squadra umana. Anche metafore come "compagno", "partner", "alter ego", "collaboratore intelligente", "amico" e "comprensione reciproca" sottolineano un alto grado di collaborazione, somiglianza e uguaglianza nei "team ibridi". Quando i partner AI consapevoli dell'uomo operano come "collaboratori umani", devono essere in grado di percepire, comprendere e reagire a un'ampia gamma di qualità comportamentali umane complesse, come l'attenzione, la motivazione, l'emozione, la creatività, la pianificazione o l'argomentazione (ad esempio Krämer et al., 2012; van den Bosch e Bronkhorst, 2018; van den Bosch et al., 2019). Pertanto, questi "partner dell'IA" o "compagni di

squadra" devono essere dotati di capacità cognitive simili a quelle umane (o umanoidi) che consentano la comprensione e la collaborazione reciproca (è questo il significato di "consapevolezza dell'essere umano").

Tuttavia, per quanto gli agenti di IA possano diventare intelligenti e autonomi sotto certi aspetti, rimarranno probabilmente macchine inconsapevoli o dispositivi speciali che supportano l'uomo in compiti specifici e complessi, almeno in un futuro prossimo. In quanto macchine digitali, sono dotate di un sistema operativo completamente diverso (digitale vs biologico) e di conseguenza di qualità e abilità cognitive diverse rispetto alle creature biologiche, come gli esseri umani e gli altri animali (Moravec, 1988; Klein et al., 2004; Korteling et al., 2018a; Shneiderman, 2020a). In generale, gli agenti digitali per il ragionamento e la risoluzione dei problemi sono paragonabili solo in modo molto superficiale alle loro controparti biologiche (ad esempio, Boden, 2017; Shneiderman, 2020b). Tenendo questo a mente, diventa sempre più importante che i professionisti umani che lavorano con sistemi avanzati di IA (ad esempio, in gruppi di lavoro militari o politici) sviluppino un tipo di ragionamento adeguato alle diverse capacità cognitive dei sistemi di IA che integrano la cognizione umana. Questo problema diventerà sempre più rilevante quando i sistemi di IA saranno più sofisticati e verrà loro lasciato un margine di autonomia maggiore. Pertanto, il presente documento cerca di fornire una maggiore chiarezza e comprensione delle caratteristiche fondamentali, delle differenze e delle idiosincrasie delle intelligenze umane/biologiche e artificiali/digitali. Nella sezione finale, viene introdotto un quadro globale per la costruzione di contenuti educativi sulla "consapevolezza dell'intelligenza". Può essere utilizzato per lo sviluppo di programmi di istruzione e formazione per gli esseri umani che dovranno utilizzare o "collaborare" con sistemi avanzati di intelligenza artificiale in un futuro prossimo e lontano.

Con l'applicazione di sistemi di intelligenza artificiale dotati di crescente autonomia, sempre più ricercatori ritengono necessario affrontare con decisione le complesse questioni reali dell'in-

telligenza di livello umano e, più in generale, dell'*intelligenza generale artificiale* (AGI [*Artificial General Intelligence*]) (ad esempio Goertzel et al., 2014). Sono già state proposte molte definizioni diverse di A(G)I (per una panoramica, si veda Russell e Norvig, 2014). Molte di esse si riducono a una *tecnologia che contiene o implica un'intelligenza (simile a quella umana)* (ad esempio, Kurzweil, 1990). Questo aspetto è problematico. La maggior parte delle definizioni utilizza il termine "intelligenza" come elemento essenziale della definizione stessa, il che rende la definizione tautologica. In secondo luogo, l'idea che l'A(G)I debba essere *simile all'uomo* sembra ingiustificata. Almeno negli ambienti naturali esistono molte altre forme e manifestazioni di comportamenti altamente complessi e intelligenti che sono molto diversi dalle specifiche capacità cognitive *umane* (per una panoramica si veda Grind, 1997). Infine, come spesso accade anche nel campo della biologia, queste definizioni di A(G)I utilizzano l'intelligenza *umana* come base o analogia centrale per ragionare sul fenomeno meno familiare dell'A(G)I (Coley e Tanner, 2012). A causa delle numerose differenze tra il substrato e l'architettura di base dell'intelligenza biologica e artificiale, questo modo *antropocentrico* di ragionare è probabilmente ingiustificato. Per questi motivi proponiamo una definizione (non antropocentrica) di "intelligenza" come "*capacità di realizzare obiettivi complessi*" (Tegmark, 2017). Questi obiettivi possono riguardare dei compiti ristretti e limitati (IA ristretta) o dei domini più ampi di compiti (AGI). Partendo da questa definizione e da una definizione di AGI proposta da Bieger et al. (2014) e da una di Grind (1997), definiamo qui l'AGI come "*capacità non biologiche di raggiungere in modo autonomo ed efficiente obiettivi complessi in un'ampia varietà di ambienti*". I sistemi AGI dovrebbero essere in grado di capire ed isolare le caratteristiche più importanti per il proprio funzionamento e apprendere in modo automatico ed efficiente diversi compiti in diversi contesti. La ricerca sull'intelligenza generale artificiale si differenzia dalla ricerca ordinaria sull'intelligenza artificiale per la versatilità e la completezza di questa intelligenza e per la ricerca

di una pratica ingegneristica che in un certo senso, come sistema, è paragonabile alla mente umana (Bieger et al., 2014).

Sarà affascinante creare copie di noi stessi in grado di apprendere più volte grazie all'interazione con i partner e diventare così in grado di collaborare sulla base di obiettivi comuni e della comprensione e dell'adattamento reciproci (ad esempio Bradshaw et al., 2012; Johnson et al., 2014). Questo sarebbe molto utile, ad esempio quando una IA con un alto grado di intelligenza sociale contribuirà a rendere più adeguate le interazioni con gli esseri umani, ad esempio nell'assistenza sanitaria o per l'intrattenimento (Wyrobek et al., 2008). Una vera collaborazione sulla base di obiettivi comuni e comprensione reciproca implica necessariamente una qualche forma di intelligenza generale umanoide. Per il momento, questo rimane un obiettivo lontano. Nel presente lavoro sosteniamo che per la maggior parte delle applicazioni potrebbe anche non essere molto pratico o necessario (e probabilmente un po' fuorviante) puntare con forza su sistemi che possiedono AGI "simili all'uomo" o abilità o qualità "simili all'uomo". Il fatto che gli esseri umani siano dotati di un'intelligenza generale non implica che le nuove forme inorganiche di intelligenza generale debbano rispettare i criteri dell'intelligenza umana. A questo proposito, il presente documento affronta il modo in cui pensiamo all'intelligenza (naturale e artificiale) in relazione al potenziale (e ai problemi reali in arrivo) dell'IA nel futuro a breve e medio termine. Ciò fornirà spunti di riflessione in previsione di un futuro difficile da prevedere per un campo dinamico come quello dell'IA.

1.2 CHE COS'È LA "VERA INTELLIGENZA"?

La premessa implicita nella nostra aspirazione di costruire sistemi AGI dotati di intelligenza umanoide è che l'intelligenza umana (generale) sia la "vera" forma di intelligenza. Questo concetto è già implicitamente declinato nel termine "Intelligenza Artificiale": è come se non fosse un'intelligenza reale, cioè reale quanto l'intelligenza non artificiale (biologica). In effetti, come esseri umani sappiamo di essere le entità con la più alta intelli-

genza mai osservata nell'Universo. E come estensione di ciò, ci piace vederci come esseri razionali in grado di risolvere un'ampia gamma di problemi complessi in ogni circostanza usando la nostra esperienza e intuizione, integrate dalle regole della logica, dell'analisi decisionale e della statistica. Non sorprende quindi che abbiamo qualche difficoltà ad accettare l'idea che potremmo essere un po' meno intelligenti di quanto continuiamo a dirci, cioè "la prossima rovina per l'umanità" (van Belkom, 2019). Ciò significa che i rapidi progressi nel campo dell'intelligenza artificiale sono accompagnati da una ricorrente ridefinizione di ciò che dovrebbe essere considerato "vera intelligenza (generale)". La concettualizzazione dell'intelligenza, cioè la capacità di raggiungere autonomamente e in modo efficiente obiettivi complessi, viene quindi continuamente adattata e ridotta a: "quelle cose che solo gli esseri umani possono fare". In linea con ciò, l'IA viene definita come "studiare come delegare ai computer i compiti in cui, al momento, le persone sono più brave" (Rich e Knight, 1991; Rich et al., 2009). Ciò include l'ideazione di soluzioni creative, l'uso flessibile di informazioni contestuali e di sfondo, l'uso dell'intuizione e del sentimento, la capacità di "pensare e capire" o il saper includere l'emozione in una considerazione (etica). Questi vengono citati come elementi propri dell'intelligenza *reale* (ad esempio, Bergstein, 2017). Ad esempio, il direttore dell'IA di Facebook e portavoce del settore, Yann LeCun, ha affermato in una *conferenza del MIT sul futuro del lavoro* che le macchine sono ancora lontane dall'avere "l'essenza dell'intelligenza". Questa include la capacità di comprendere il mondo fisico abbastanza bene da fare previsioni su aspetti fondamentali di esso: osservare una cosa e poi usare le conoscenze di base per capire quali altre cose devono essere vere. Un altro modo per dirlo è che le macchine non hanno il *senso comune* (Bergstein, 2017), ad esempio i sottomarini, che non sanno nuotare (van Belkom, 2019). Quando le capacità umane esclusive diventano gli unici punti cardinali nella nostra navigazione, rischiamo di perdere di vista alcuni problemi significativi che invece avrebbero la priorità.

Per chiarire questo punto, forniremo innanzitutto alcune informazioni sulla natura di base dell'intelligenza umana e artificiale. Ciò è necessario per la realizzazione di un'adeguata consapevolezza dell'intelligenza (*Intelligence Awareness*) e di un'adeguata ricerca e formazione che anticipino lo sviluppo e l'applicazione dell'I(G)A. Per il momento abbiamo tre concetti essenziali che possono (e devono) essere ulteriormente elaborati in futuro.

- Per quanto riguarda i compiti cognitivi, probabilmente siamo meno intelligenti di quanto pensiamo. Allora perché dovremmo concentrarci su un'intelligenza artificiale *simile a quella umana*?
- Sono possibili molte forme diverse di intelligenza e l'intelligenza generale non coincide necessariamente con l'intelligenza generale *umanoide* (o "AGI a livello umano").
- L'intelligenza artificiale spesso non è necessaria; molti problemi complessi possono essere affrontati in modo efficace anche utilizzando più IA ristrette. 1

2. PROBABILMENTE NON SIAMO COSÌ INTELLIGENTI COME PENSIAMO

Quanto siamo intelligenti? La risposta a questa domanda è determinata in larga misura dalla prospettiva da cui si considera la questione, e quindi dalle misure e dai criteri di intelligenza che si scelgono. Per esempio, potremmo confrontare la natura e le capacità dell'intelligenza umana con quella di altre specie animali. In questo caso sembriamo molto intelligenti. Grazie alla nostra enorme capacità di apprendimento, abbiamo di gran lunga il più vasto armamentario di abilità cognitive [2] per risolvere autonomamente problemi complessi e raggiungere obiettivi complessi. In questo modo possiamo risolvere un'enorme varietà di problemi aritmetici, concettuali, spaziali, economici, socio-organizzativi, politici, ecc. I primati, che si differenziano solo leggermente da noi in termini genetici, sono molto indie-

tro in questo senso. Possiamo quindi legittimamente qualificare l'uomo, rispetto alle altre specie animali che conosciamo, come altamente intelligente.

2.1 CAPACITÀ COGNITIVA LIMITATA

Tuttavia, possiamo anche guardare oltre questa "prospettiva *relativa* interspecie" e cercare di qualificare la nostra intelligenza in termini più *assoluti*, cioè utilizzando una scala che va da zero a ciò che è fisicamente possibile. Ad esempio, potremmo considerare la capacità computazionale di un cervello umano come un sistema fisico (Bostrom, 2014; Tegmark, 2017). L'idea prevalente a questo proposito tra gli scienziati dell'IA è che l'intelligenza sia in ultima analisi una questione di informazione e di calcolo, e (quindi) non di carne, sangue e atomi di carbonio. In linea di principio, non esiste alcuna legge fisica che impedisca che possano essere costruiti sistemi fisici (costituiti da quark e atomi, come il nostro cervello) con una potenza di calcolo e un'intelligenza molto superiori a quelle del cervello umano. Ciò sembra implicare che non esiste una legge fisica insormontabile per cui le macchine non possano un giorno diventare molto più intelligenti di noi sotto tutti i punti di vista possibili (Tegmark, 2017). La nostra intelligenza è quindi *relativamente* elevata rispetto a quella di altri animali, ma in termini assoluti potrebbe essere molto limitata nella sua capacità di calcolo fisico, anche se solo per le dimensioni limitate del nostro cervello e il suo numero massimo possibile di neuroni e cellule della glia (ad esempio, Kahle, 1979).

Per definire e valutare ulteriormente la nostra intelligenza (biologica), possiamo anche discutere l'evoluzione e la natura delle nostre capacità di pensiero biologico. In quanto rete neurale biologica di carne e sangue, necessaria per la sopravvivenza, il nostro cervello ha subito un processo di ottimizzazione evolutiva di oltre un miliardo di anni. In questo lungo periodo, si è sviluppato in un sistema altamente efficace ed efficiente per regolare le funzioni biologiche essenziali e svolgere compiti percettivo-motori e di riconoscimento dei modelli, come la raccolta del cibo, la lotta, la fuga e l'accoppiamento. Quasi durante tutta la nostra

evoluzione, le reti neurali del nostro cervello sono state ulteriormente ottimizzate per questi processi motori biologici e percettivi di base, che sono anche l'origine delle nostre abilità pratiche quotidiane come cucinare, fare giardinaggio o i lavori domestici. Forse a causa della nostra competenza in questo tipo di compiti dimentichiamo che sono caratterizzati da una complessità *computazionale* estremamente elevata (ad esempio, Moravec, 1988). Ad esempio, quando allacciamo le scarpe, molti milioni di segnali entrano ed escono attraverso un gran numero di sistemi sensoriali diversi, dai corpi tendinei e dai fusi muscolari delle estremità alla retina, agli organi otolitici e ai canali semicircolari della testa (ad esempio, Brodal, 1981). Questa enorme quantità di informazioni provenienti da molti sistemi percettivo-motori diversi viene elaborata di continuo nelle reti neurali del nostro cervello, senza sforzo e persino senza attenzione cosciente (Minsky, 1986; Moravec, 1988; Grind, 1997). Per raggiungere questo obiettivo, il cervello dispone di una serie di meccanismi di lavoro universali (intrinseci), come l'associazione e l'apprendimento associativo (Shatz, 1992; Bar, 2007), il potenziamento e la facilitazione (Katz e Miledi, 1968; Bao et al., 1997), la saturazione e l'inibizione laterale (Isaacson e Scanziani, 2011; Korteling et al., 2018a).

Queste capacità biologiche e percettivo-motorie di base si sono sviluppate e consolidate nel corso di molti milioni di anni. Molto più tardi nella nostra evoluzione - in realtà solo molto recentemente - hanno iniziato a svilupparsi le nostre capacità cognitive e le nostre funzioni razionali. Queste abilità o capacità cognitive hanno probabilmente meno di 100 mila anni, il che può essere definito a uno stadio "embrionale" nella scala temporale dell'evoluzione (ad esempio, Petraglia e Korisettar, 1998; McBrearty e Brooks, 2000; Henshilwood e Marean, 2003). Inoltre, questa ristrettissima fase di conquiste umane è stata costruita su queste "antiche" intelligenze neurali per funzioni essenziali di sopravvivenza. Quindi, le nostre capacità cognitive "superiori" si sono sviluppate *da* e *con* questi meccanismi di regolazione (neuro)biologica (Damasio, 1994; Korteling e Toet, 2020). Di conseguenza, non dovrebbe sorprendere che le capacità del nostro

cervello in queste recenti funzioni cognitive siano ancora piuttosto limitate. Queste limitazioni si manifestano in molti modi diversi, ad esempio:

- La quantità di informazioni cognitive che possiamo elaborare consapevolmente (la nostra memoria di lavoro, span o attenzione) è molto limitata (Simon, 1955). La capacità della nostra memoria di lavoro è di circa 10-50 bit al secondo (Tegmark, 2017).
- La maggior parte dei compiti cognitivi, come la lettura di un testo o i calcoli, richiedono la nostra piena attenzione e di solito abbiamo bisogno di molto tempo per eseguirli. Le calcolatrici mobili possono eseguire calcoli milioni di volte più complessi di noi (Tegmark, 2017).
- Sebbene possiamo elaborare molte informazioni in parallelo, non possiamo eseguire simultaneamente compiti cognitivi che richiedono deliberazione e attenzione, cioè il "multitasking" (Korteling, 1994; Rogers e Monsell, 1995; Rubinstein, Meyer e Evans, 2001).
- Le conoscenze e le abilità cognitive acquisite dalle persone (memoria) tendono a decadere nel tempo, molto più delle abilità percettivo-motorie. A causa di questa limitata ritenzione delle informazioni, dimentichiamo facilmente porzioni sostanziali di ciò che abbiamo appreso (Wingfield e Byrnes, 1981).

2.2 DISTORSIONI COGNITIVE RADICATE

La nostra limitata capacità di elaborazione dei compiti cognitivi non è l'unico fattore che determina la nostra intelligenza cognitiva. Oltre a una capacità di elaborazione complessivamente limitata, l'elaborazione dell'informazione cognitiva umana presenta distorsioni sistematiche. Queste si manifestano in molti pregiudizi cognitivi (Tversky e Kahneman, 1973, Tversky e Kahneman, 1974). Le distorsioni cognitive sono tendenze, inclinazioni o disposizioni sistematiche e universali che alterano o distorcono i processi informativi in modo tale da rendere il

loro risultato impreciso, non ottimale o semplicemente sbagliato (ad esempio Lichtenstein e Slovic, 1971; Tversky e Kahneman, 1981). Molte distorsioni si manifestano quasi allo stesso modo in diverse situazioni decisionali (Shafir e LeBoeuf, 2002; Kahneman, 2011; Toet et al., 2016). La letteratura fornisce descrizioni e dimostrazioni di oltre 200 distorsioni. Queste tendenze sono in gran parte implicite e inconsce e si percepiscono in modo del tutto naturale e autoevidente quando siamo consapevoli di queste inclinazioni cognitive (Pronin et al., 2002; Risen, 2015; Korteling et al., 2018b). Per questo motivo vengono spesso definite "intuitive" (Kahneman e Klein, 2009) o "irrazionali" (Shafir e LeBoeuf, 2002). Il ragionamento distorto può portare a risultati abbastanza accettabili in situazioni naturali o quotidiane, soprattutto se si tiene conto del tempo necessario per il ragionamento (Simon, 1955; Gigerenzer e Gaissmaier, 2011). Tuttavia, le persone spesso deviano dalla razionalità e/o dai principi della logica, del calcolo e della probabilità in modi sconsigliabili (Tversky e Kahneman, 1974; Shafir e LeBoeuf, 2002), portando a decisioni non ottimali in termini di tempo e sforzi investiti (costi), date le informazioni disponibili e i benefici attesi.

Le distorsioni sono in gran parte causate da caratteristiche e meccanismi *intrinseci* (o strutturali) del cervello come rete neurale (Korteling et al., 2018a; Korteling e Toet, 2020). Fondamentalmente, questi meccanismi - come l'associazione, la facilitazione, l'adattamento o l'inibizione laterale - comportano una modifica dei dati originali o disponibili e della loro elaborazione (ad esempio, la ponderazione della loro importanza). Per esempio, l'inibizione laterale è un processo neurale universale che porta ad aumentare le differenze nell'attività neurale (potenziamento del contrasto), molto utile per le funzioni percettivo-motorie, il mantenimento dell'integrità fisica e l'allostasi (cioè le funzioni di sopravvivenza biologica). Per queste funzioni il nostro sistema nervoso è stato ottimizzato per milioni di anni. Tuttavia, le funzioni cognitive "superiori", come il pensiero concettuale, il ragionamento probabilistico o il calcolo, sono state sviluppate solo molto recentemente nel corso dell'evoluzione. Queste funzioni

hanno probabilmente meno di 100 mila anni e possono quindi essere definite "embrionali" nella scala temporale dell'evoluzione (ad esempio, McBrearty e Brooks, 2000; Henshilwood e Marean, 2003; Petraglia e Korisettar, 2003). Inoltre, l'evoluzione non ha potuto sviluppare queste nuove funzioni cognitive da zero, ma ha dovuto costruire queste conquiste umane embrionali a partire da un "antico" patrimonio neurale che serviva per le funzioni biologiche essenziali di sopravvivenza (Moravec, 1988). Poiché le funzioni cognitive richiedono in genere un calcolo esatto e un'adeguata ponderazione dei dati, le trasformazioni dei dati, come l'inibizione laterale, possono facilmente portare a distorsioni sistematiche (cioè a pregiudizi) nell'elaborazione delle informazioni cognitive. Alcuni esempi del gran numero di distorsioni causate dalle proprietà intrinseche delle reti neurali biologiche sono: *Anchoring bias* (che orienta le decisioni verso informazioni acquisite in precedenza, Furnham e Boo, 2011; Tversky e Kahneman, 1973, Tversky e Kahneman, 1974), l'*Hindsight bias* (la tendenza a percepire erroneamente gli eventi come inevitabili o più probabili una volta che si sono verificati, Hoffrage et al, 2000; Roese e Vohs, 2012), l'*availability bias* (giudicare la frequenza, l'importanza o la probabilità di un evento in base alla facilità con cui si ricordano i casi rilevanti, Tversky e Kahnemann, 1973; Tversky e Kahneman, 1974) e il *confirmation bias* (la tendenza a selezionare, interpretare e ricordare le informazioni in modo da confermare le proprie preconoscenze, opinioni e aspettative, Nickerson, 1998). Oltre a queste limitazioni (strutturali) intrinseche delle reti neurali (biologiche), i bias possono anche avere origine da principi evolutivi funzionali che promuovono la sopravvivenza dei nostri antenati che, come cacciatori-raccoglitori, vivevano in piccoli gruppi affiatati (Haselton et al., 2005; Tooby e Cosmides, 2005). I pregiudizi cognitivi [sinonimo di "distorsioni cognitive" o "bias cognitivi"] possono essere causati da una mancata corrispondenza tra le "euristiche" razionalizzate evolutivamente ("razionalità evolutiva": Haselton et al., 2009) e il contesto o l'ambiente attuale (Tooby e Cosmides, 2005). Secondo questa visione, le stesse euristiche che hanno ottimizzato le possibilità

di sopravvivenza dei nostri antenati nel loro ambiente (naturale) possono portare a comportamenti disadattivi (distorti) quando vengono utilizzate nei nostri contesti attuali (artificiali). I bias che sono stati considerati come esempi di questo tipo di disadattamento sono l'*Action bias* (preferire l'azione anche quando non c'è una giustificazione razionale per farlo, Baron e Ritov, 2004; Patt e Zeckhauser, 2000), la prova sociale (la tendenza a rispecchiare o copiare le azioni e le opinioni degli altri, Cialdini, 1984), la "Tragedia dei beni comuni" (privilegiare gli interessi personali rispetto al bene comune della comunità, Hardin, 1968) e l'*Ingroup bias* (favorire il proprio gruppo rispetto a quello degli altri, Taylor e Doria, 1981).

Questo carattere radicato (neuralmente o evolutivamente intrinseco) del pensiero distorto rende improbabile che metodi semplici e diretti come interventi di formazione o corsi di sensibilizzazione siano efficaci per attenuare i pregiudizi. Questa difficoltà nell'attenuare i pregiudizi sembra effettivamente supportata dalla letteratura (Korteling et al., 2021).

3. L'INTELLIGENZA GENERALE NON COINCIDE CON L'INTELLIGENZA DI TIPO UMANO

3.1 DIFFERENZE FONDAMENTALI TRA INTELLIGENZA BIOLOGICA E ARTIFICIALE

Spesso pensiamo all'intelligenza e prendiamo decisioni su di essa mantenendo come riferimento ovvio e non-ambiguo una concezione antropocentrica della nostra intelligenza. Tendiamo a usare questa concezione come base per ragionare su altri fenomeni di intelligenza a noi meno familiari, come altre forme di intelligenza biologica e artificiale (Coley e Tanner, 2012). Questo può portare a domande e idee affascinanti. Un esempio è la discussione su come e quando si raggiungerà il punto di "intelligenza di livello umano". Ad esempio, Ackermann. (2018) scrive: "Prima di raggiungere la superintelligenza, l'IA generale

implica che una macchina avrà le stesse capacità cognitive di un essere umano". I ricercatori si interrogano ampiamente sul momento in cui raggiungeremo l'IA generale (ad esempio, Goertzel, 2007; Müller e Bostrom, 2016). Riteniamo che questo tipo di domande non siano del tutto pertinenti. Ci sono (in linea di principio) molti tipi diversi di intelligenza (generale) concepibili, e l'intelligenza simile a quella umana è solo uno di quelli. Ciò significa, ad esempio, che lo sviluppo dell'intelligenza artificiale è determinato dai vincoli della fisica e della tecnologia, e non da quelli dell'evoluzione biologica. Quindi, allo stesso modo in cui l'intelligenza di un ipotetico extraterrestre che visita la Terra è probabile che abbia una struttura (in-)organica diversa, con caratteristiche, punti di forza e debolezze differenti dagli umani, questo vale anche per le forme artificiali di intelligenza (generale). Di seguito riassumiamo brevemente alcune differenze fondamentali tra intelligenza umana e artificiale (Bostrom, 2014):

- Struttura di base: l'intelligenza biologica (al carbonio) si basa su un "*wetware*" neurale che è fondamentalmente diverso dall'intelligenza artificiale (al silicio). A differenza del wetware biologico, nei sistemi al silicio, o digitali, "hardware" e "software" sono indipendenti l'uno dall'altro (Kosslyn e Koenig, 1992). Quando un sistema biologico ha imparato una nuova abilità, questa sarà limitata al sistema stesso. Al contrario, se un sistema di intelligenza artificiale ha appreso una certa abilità, gli algoritmi che la costituiscono possono essere copiati direttamente in tutti gli altri sistemi digitali simili.
- Velocità: i segnali dei sistemi di intelligenza artificiale si propagano quasi alla velocità della luce. Negli esseri umani, la velocità di conduzione dei nervi procede con una velocità massima di 120 m/s, che è estremamente lenta nella scala temporale dei computer (Siegel e Sapru, 2005).
- Connettività e comunicazione: le persone non possono comunicare direttamente tra loro. Comunicano attraverso il linguaggio e i gesti con una larghezza di banda limitata. Si

tratta di una comunicazione più lenta e difficile rispetto a quella dei sistemi di intelligenza artificiale, che possono essere direttamente collegati tra loro. Grazie a questa connessione diretta, possono anche collaborare sulla base di algoritmi integrati.

- Aggiornabilità e scalabilità: i sistemi di intelligenza artificiale non hanno quasi alcun vincolo per quanto riguarda l'aggiornamento o la scalabilità e/o riconfigurazione, in modo da avere gli algoritmi giusti e le capacità di elaborazione e archiviazione dei dati necessarie per i compiti che devono svolgere. Questa capacità di espansione rapida e strutturale e di miglioramento immediato difficilmente si applica alle persone.

Al contrario, la biologia fa molto con poco: i cervelli organici sono milioni di volte più efficienti nel consumo di energia rispetto ai computer. Il cervello umano consuma meno energia di una lampadina, mentre un supercomputer con prestazioni di calcolo paragonabili consuma elettricità sufficiente ad alimentare un villaggio (Fischetti, 2011).

Queste differenze nella struttura di base, nella velocità, nella connettività, nell'aggiornabilità, nella scalabilità e nel consumo di energia porteranno necessariamente a qualità e limiti diversi tra l'intelligenza umana e quella artificiale. La nostra velocità di risposta a stimoli semplici è, ad esempio, molte migliaia di volte più lenta di quella dei sistemi artificiali. I sistemi informatici possono molto facilmente essere collegati direttamente tra loro e, in quanto tali, possono far parte di un unico sistema integrato. Ciò significa che i sistemi di IA non devono essere visti come entità individuali che possono facilmente lavorare fianco a fianco o avere incomprensioni reciproche. Inoltre, se due sistemi di intelligenza artificiale sono impegnati in un compito, il rischio di commettere errori di comunicazione è minimo (si pensi ai veicoli autonomi che si avvicinano a un incrocio). Dopo tutto, sono parti intrinsecamente connesse dello stesso sistema e dello stesso algoritmo (Gerla et al., 2014).

3.2 COMPLESSITÀ E PARADOSSO DI MORAVEC

Poiché i cervelli biologici, basati sul carbonio, e i computer digitali, basati sul silicio, sono ottimizzati per tipi di compiti completamente diversi (ad esempio, Moravec, 1988; Korteling et al., 2018b), l'intelligenza umana e quella artificiale presentano differenze fondamentali e probabilmente di ampia portata. A causa di queste differenze, può essere molto fuorviante utilizzare la nostra mente come base, modello o analogia per ragionare sull'intelligenza artificiale. Ciò può portare a concezioni errate, ad esempio sulle presunte capacità degli esseri umani e delle IA di svolgere compiti complessi. I difetti che ne derivano riguardo alle capacità di elaborazione delle informazioni emergono spesso nella letteratura psicologica, in cui "complessità" e "difficoltà" sono termini usati in modo intercambiabile (si vedano ad esempio Wood et al., 1987; McDowd e Craik, 1988). La complessità di un compito viene poi valutata in modo antropocentrico, cioè in base al grado di padronanza del compito da parte degli esseri umani. Quindi, utilizziamo la *difficoltà* nell'eseguire o padroneggiare un compito come misura della sua *complessità*, e le prestazioni del compito (velocità, errori) come misura dell'abilità e dell'intelligenza di chi lo esegue. Sebbene questo possa essere talvolta accettabile nella ricerca psicologica, potrebbe essere fuorviante se ci sforziamo di comprendere l'intelligenza dei sistemi di intelligenza artificiale. Per noi è molto più difficile moltiplicare due numeri casuali di sei cifre che riconoscere un amico in una fotografia. Ma quando si tratta di operazioni di conteggio o aritmetiche, i computer sono migliaia di volte più veloci e più bravi, mentre gli stessi sistemi hanno fatto passi avanti solo di recente nel riconoscimento delle immagini (che è riuscito solo quando è stata sviluppata la tecnologia del deep learning, basata su alcuni principi delle reti neurali biologiche). In generale: i compiti cognitivi che sono relativamente difficili per il cervello umano (e che quindi troviamo soggettivamente difficili) non devono essere necessariamente complessi dal punto di vista computazionale (ad esempio, in termini di operazioni

aritmetiche, logiche e astratte oggettive). E viceversa: i compiti che sono relativamente facili per il cervello (riconoscimento di schemi, compiti percettivo-motori, compiti ben allenati) non devono necessariamente essere computazionalmente semplici. Questo fenomeno è stato definito il *paradosso di Moravec*: quello che è facile per l'antica "tecnologia" neurale delle persone e difficile per la moderna tecnologia digitale dei computer (e viceversa). Hans Moravec (1988) ha scritto: "È relativamente facile far sì che i computer mostrino prestazioni di livello adulto nei test di intelligenza o nel gioco della dama, ed è difficile o impossibile dar loro le capacità di un bambino di un anno quando si tratta di percezione e mobilità".

3.3 INTELLIGENZA PERCETTIVO-MOTORIA SUPERIORE DELL'UOMO

Il paradosso di Moravec implica che le reti neurali biologiche sono intelligenti in modi diversi rispetto alle reti neurali artificiali. L'intelligenza non si limita ai problemi o agli obiettivi che noi esseri umani, dotati di intelligenza biologica, troviamo difficili (Grind, 1997). L'intelligenza, definita come la capacità di realizzare obiettivi complessi o di risolvere problemi complessi, è molto di più. Secondo Moravec (1988), il ragionamento ad alto livello richiede pochissimo calcolo, ma le abilità percettivo-motorie di basso livello richiedono enormi risorse computazionali. Se esprimiamo la complessità di un problema in termini di numero di calcoli elementari necessari per risolverlo, allora la nostra intelligenza percettivo-motoria biologica è *altamente superiore* alla nostra intelligenza cognitiva. La nostra intelligenza organica percettivo-motoria è particolarmente brava nell'elaborazione associativa di invarianti di ordine superiore ("modelli") nell'informazione ambientale. Questi sono computazionalmente più complessi e contengono più informazioni dei singoli elementi semplici (Gibson, 1966, Gibson, 1979). Un esempio delle nostre capacità percettivo-motorie superiori è l'*effetto* di *superiorità dell'oggetto*: percepiamo e interpretiamo gli oggetti interi in modo più rapido ed efficace rispetto ai singoli elementi (più semplici) che li compongono (Weisstein e Harris, 1974; McClelland,

1978; Williams e Weisstein, 1978; Pomerantz, 1981). Pertanto, anche le lettere vengono percepite in modo più accurato quando vengono presentate come parte di una parola rispetto a quando vengono presentate isolatamente, ovvero l'effetto di superiorità della parola (ad esempio, Reicher, 1969; Wheeler, 1970). Quindi, la *difficoltà* di un compito non indica necessariamente la sua *complessità* intrinseca. Come dice Moravec (1988): "Siamo tutti olimpionici prodigiosi nelle aree percettive e motorie, così bravi da far sembrare facile il difficile". Il pensiero astratto, però, è un trucco nuovo, che ha forse meno di 100 mila anni. Non l'abbiamo ancora imparato. Non è così intrinsecamente difficile; lo sembra solo quando lo facciamo".

3.4 L'IPOTESI DI UN'INTELLIGENZA ARTIFICIALE SIMILE A QUELLA UMANA

Quindi, se dovessero esistere sistemi di IA con intelligenza generale che possono essere utilizzati per un'ampia gamma di problemi e obiettivi complessi, queste macchine AGI avrebbero probabilmente un profilo di intelligenza completamente diverso rispetto a quello degli esseri umani, oltre a presentare altre qualità cognitive (Goertzel, 2007). Ciò avverrà anche se riusciremo a costruire agenti di IA che mostrino un comportamento simile al nostro e se saranno in grado di adattarsi al nostro modo di pensare e di risolvere i problemi, al fine di promuovere la collaborazione tra uomo e IA. A meno che non decidiamo di *degradare* deliberatamente le capacità dei sistemi di IA (il che non sarebbe molto intelligente), le capacità e le abilità sottostanti dell'uomo e delle macchine per quanto riguarda la raccolta e l'elaborazione delle informazioni, l'analisi dei dati, il ragionamento probabilistico, la logica, la capacità di memoria ecc. rimarranno comunque dissimili. A causa di queste differenze, dovremmo concentrarci su sistemi che ci *completino* efficacemente e che rendano il sistema uomo-IA più forte ed efficace. Invece di perseguire un'IA di livello umano, sarebbe più utile concentrarsi su macchine autonome e sistemi (di supporto) che colmino o estendano le

molteplici lacune dell'intelligenza cognitiva umana. Ad esempio, mentre le persone sono costrette, a causa della lentezza e di altre limitazioni dei cervelli biologici, a pensare euristicamente in termini di obiettivi, virtù, regole e norme espresse in un linguaggio (informale), l'IA ha già stabilito eccellenti capacità di elaborare e calcolare direttamente su dati altamente complessi. Pertanto, per l'esecuzione di compiti cognitivi specifici (ristretti) (logici, analitici, computazionali), la moderna intelligenza digitale può essere più efficace ed efficiente dell'intelligenza biologica. L'IA può quindi contribuire a produrre risposte migliori per problemi complessi utilizzando elevate quantità di dati, insiemi coerenti di principi e obiettivi etici, ragionamenti probabilistici e logici (ad esempio, Korteling et al., 2018b). Quindi, ipotizziamo che, in ultima analisi, lo sviluppo di sistemi di IA a supporto del processo decisionale umano possa apparire come il modo più efficace per compiere scelte migliori o sviluppare soluzioni migliori per questioni complesse. Perciò, la cooperazione e la divisione dei compiti tra persone e sistemi di IA dovrà essere determinata principalmente dalle loro qualità specifiche. Ad esempio, i compiti o le parti dei compiti che fanno appello a capacità in cui i sistemi di IA eccellono dovranno essere gestiti in misura minore dalle persone, per cui probabilmente sarà necessaria meno formazione. I sistemi di IA sono già molto più bravi delle persone nel raccogliere (selezionare) ed elaborare (pesare, dare priorità, analizzare, combinare) grandi quantità di dati in modo logicamente e aritmeticamente corretto. Lo fanno in modo rapido, preciso e affidabile. Sono anche più stabili (coerenti) degli esseri umani, non hanno stress ed emozioni e hanno una grande perseveranza e una conservazione delle conoscenze e delle abilità molto migliore rispetto alle persone. In quanto macchine, servono le persone in modo completo e senza alcun interesse personale o secondo fine. Sulla base di queste qualità, i sistemi di IA possono effettivamente sollevare le persone da compiti o parti di compiti. Tuttavia, è importante che le persone continuino a padroneggiare tali compiti fino a un certo punto, in modo da poterli rilevare

o intervenire adeguatamente in caso di fallimento del sistema automatico.

In generale, le persone sono più adatte dei sistemi di IA per uno spettro molto più ampio di compiti cognitivi e sociali in un'ampia varietà di circostanze ed eventi (imprevisti) (Korteling et al., 2018b). Per il momento, le persone sono anche più brave nell'interazione sociale-psicosociale. Ad esempio, è difficile per i sistemi di intelligenza artificiale interpretare il linguaggio e il simbolismo umano. Ciò richiede un quadro di riferimento molto ampio che, almeno fino ad ora e per il prossimo futuro, è difficile da raggiungere nell'ambito dell'IA. Come risultato di tutte queste differenze, le persone sono ancora più brave a rispondere (come team flessibile) a situazioni inaspettate e imprevedibili e a escogitare in modo creativo possibilità e soluzioni in compiti aperti e mal definiti e in molte circostanze diverse e magari inaspettate. Le persone dovranno fare un uso supplementare delle loro qualità umane specifiche (cioè ciò che le persone sono relativamente brave a fare) e allenarsi sulle competenze pertinenti. Inoltre, i membri del team umano dovranno imparare a gestire bene i limiti generali delle IA. Con un'adeguata divisione dei compiti, che sfrutti le qualità e i limiti specifici degli esseri umani e dei sistemi di IA, è possibile aggirare i pregiudizi decisionali umani e ci si possono aspettare prestazioni migliori. Ciò significa che il potenziamento di un team con macchine intelligenti che hanno meno vincoli e pregiudizi cognitivi può avere un valore aggiunto maggiore rispetto alla collaborazione tra esseri umani e IA che hanno sviluppato gli stessi pregiudizi (umani). Sebbene la cooperazione in team con sistemi di IA possa richiedere un addestramento supplementare per gestire efficacemente questi vari pregiudizi, l'eterogeneità è probabilmente migliore e più sicura. Questo apre anche la possibilità di una combinazione tra un alto livello di controllo umano significativo e alti livelli di automazione, che probabilmente produrrà i sistemi umani-IA più efficaci e sicuri (Elands et al., 2019; Shneiderman, 2020a). In breve: l'intelligenza umana non è l'ideale a cui tendere per l'intelligenza generale; invece di puntare a un'intelligenza artificiale

simile a quella umana, il perseguimento dell'intelligenza artificiale dovrebbe quindi concentrarsi su un'efficace *intelligenza artificiale digitale/di silicio in combinazione* con una capacità di prendere in carico e svolgere dei compiti.

3.5 SPIEGABILITÀ E FIDUCIA

Gli sviluppi relativi all'apprendimento artificiale, o all'apprendimento profondo (di rinforzo), in particolare, sono stati rivoluzionari. L'apprendimento profondo simula una rete che assomiglia alle reti neurali stratificate del nostro cervello. Basandosi su grandi quantità di dati, la rete impara a riconoscere schemi e collegamenti con un alto livello di precisione e poi li collega a una serie di azioni senza conoscere i nessi causali sottostanti. Ciò implica che è difficile fornire all'intelligenza artificiale dell'apprendimento profondo una sorta di trasparenza su come o perché ha fatto una determinata scelta, ad esempio esprimendo un ragionamento intelligibile (per gli esseri umani) sul suo processo decisionale, come facciamo noi (ad esempio, Belkom, 2019). Inoltre, ragionare sulle decisioni come fanno gli esseri umani è un processo molto variabile e che non avviene sempre (almeno negli esseri umani). In genere gli esseri umani non sono consapevoli delle loro cognizioni o atteggiamenti impliciti e quindi non sono in grado di parlarne adeguatamente. Per molti esseri umani è quindi piuttosto difficile fare introspezione sui propri stati mentali, nella misura in cui questi sono consapevoli, e collegare i risultati di questa analisi a etichette e descrizioni verbali (ad esempio, Nosek et al. (2011)). In primo luogo, il cervello umano difficilmente rivela come crea pensieri coscienti (ad esempio, Feldman-Barret, 2017). In realtà, ci dà l'illusione che i risultati rivelino un funzionamento interno. In altre parole, i nostri pensieri coscienti non ci dicono nulla sul modo in cui sono nati. Non esiste nemmeno un indicatore soggettivo che distingua i processi di ragionamento corretti da quelli errati (Kahneman e Klein, 2009). Il decisore non ha quindi modo di distinguere tra i pensieri corretti, che derivano da una conoscenza e da un'espe-

rienza autentiche, e quelli errati, che derivano da processi neuroevolutivi, tendenze e intuizioni primarie inappropriate. Quindi potremmo chiederci: non è più affidabile avere una vera scatola nera, piuttosto che ascoltarne una che fa speculazioni? Inoltre, secondo Werkhoven et al. (2018) la richiesta di spiegabilità, osservabilità o trasparenza (Belkom, 2019; van den Bosch et al., 2019) può indurre i sistemi intelligenti artificiali a limitare il loro potenziale beneficio per la società umana a ciò che può essere compreso dagli esseri umani.

Naturalmente non dobbiamo fidarci ciecamente dei risultati generati dall'IA. Come in altri campi della tecnologia complessa (ad esempio, la modellazione e la simulazione), i sistemi di IA devono essere verificati (rispettando le specifiche) e convalidati (raggiungendo gli obiettivi del sistema) in relazione agli obiettivi per i quali il sistema è stato progettato. In generale, quando un sistema è adeguatamente verificato e validato, può essere considerato sicuro, protetto e adatto allo scopo. Merita quindi la nostra fiducia per ragioni (logicamente) comprensibili e oggettive (anche se gli errori possono ancora verificarsi). Allo stesso modo, la gente si fida delle prestazioni degli aerei e dei telefoni cellulari, nonostante la quasi totale ignoranza dei loro complessi processi interni. Come il nostro cervello, le reti neurali artificiali sono fondamentalmente non-trasparenti (Nosek et al., 2011; Feldman-Barret, 2017). Per tal motivo la fiducia nell'IA dovrebbe basarsi principalmente sulle sue prestazioni oggettive. Questa è una base più importante rispetto a quella che fornisce fiducia sulla base di impressioni, storie o immagini soggettive (ingannevoli) volte a convincere e attrarre l'utente. Sulla base di ricerche di validazione empirica, gli sviluppatori e gli utenti possono verificare con i propri occhi quanto il sistema si stia comportando bene rispetto all'insieme di valori e obiettivi per cui la macchina è stata progettata. A un certo punto, gli esseri umani possono desiderare di raggiungere gli obiettivi a costi inferiori e con risultati migliori, e ciò accade quando accettiamo alcune soluzioni, benché siano meno trasparenti di altre (Werkhoven et al., 2018).

4. L'IMPATTO DELLA TECNOLOGIA DELL'IA RISTRETTA E MULTIPLA

4.1 L'INTELLIGENZA ARTIFICIALE COME SANTO GRAAL

L'AGI, come l'intelligenza generale umana, avrebbe molti vantaggi evidenti rispetto all'IA ristretta (limitata, debole, specializzata). Un sistema AGI sarebbe molto più flessibile e adattabile. Sulla base di processi di addestramento e ragionamento generici, capirebbe autonomamente come risolvere vari problemi in tutti i tipi di domini diversi in relazione al loro contesto (ad esempio, Kurzweil, 2005). I sistemi AGI richiedono anche un numero molto inferiore di interventi umani per sistemare i problemi generati da elementi parziali, sfaccettature e prospettive in situazioni complesse. L'AGI comprenderebbe davvero i problemi e sarebbe in grado di vederli da diverse prospettive (come possono fare teoricamente anche le persone). Una caratteristica degli attuali strumenti di IA (ristretti) è che sono abili in un compito molto specifico, in cui spesso riescono a raggiungere livelli superiori a quelli umani (ad esempio Goertzel, 2007; Silver et al., 2017). Questi compiti specifici sono stati ben definiti e strutturati. I sistemi di intelligenza artificiale ristretta sono meno adatti, o del tutto inadatti, a compiti o ambienti di lavoro che offrono poca struttura, coerenza, regole o indicazioni, in cui può verificarsi ogni sorta di evento imprevisto, raro o non comune (ad esempio, emergenze). Conoscere e seguire procedure fisse di solito non porta a soluzioni adeguate in queste circostanze variabili. Nel contesto di cambiamenti (imprevisti) negli obiettivi o nelle circostanze, l'adeguatezza dell'IA attuale è notevolmente ridotta perché non può ragionare da una prospettiva generale e adattarsi di conseguenza (Lake et al., 2017; Horowitz, 2018). Come per i sistemi di IA ristretti, è necessario che le persone supervisionino queste deviazioni per consentire prestazioni flessibili e adattive del sistema. Pertanto, la ricerca dell'IA può essere considerata come la ricerca di una sorta di Santo Graal.

4.2 L'INTELLIGENZA ARTIFICIALE MULTIPLA È LA PIÙ IMPORTANTE ORA!

Le prospettive potenzialmente alte dell'IA, tuttavia, non implicano che l'IA sarà il fattore più importante della futura ricerca e sviluppo sull'IA, almeno nel breve e medio termine. Quando riflettiamo sui grandi e potenziali benefici dell'intelligenza generale tendiamo a considerare le applicazioni di IA ristrette come entità separate che possono benissimo essere superate da un'IA più ampia che presumibilmente può occuparsi di tutto. Ma proprio come il nostro mondo moderno si è evoluto rapidamente grazie a una serie di innovazioni tecnologiche specifiche (limitate), così anche l'ampia gamma di applicazioni di IA emergenti avrà un impatto tecnologico e sociale rivoluzionario (Peeters et al., 2020). Ciò sarà ancora più rilevante nel futuro mondo dei big data, in cui tutto è connesso a tutto attraverso l'*Internet delle cose*. Sarà quindi molto più redditizio e vantaggioso sviluppare e costruire varianti di IA (non simili all'uomo) che eccellano in aree in cui l'uomo è di per sé limitato. Non è azzardato ipotizzare che le molteplici varianti di applicazioni di IA ristrette possano gradualmente diventare più interconnesse. In questo modo, ci si può aspettare uno sviluppo verso un insieme sempre più ampio di applicazioni di IA integrate. Inoltre, è già possibile addestrare un modello linguistico di IA (*Generative Pre-trained Transformer3*, GPT-3) con un gigantesco set di dati e poi fargli apprendere vari compiti sulla base di pochi esempi - un apprendimento in pochi passaggi. GPT-3 (sviluppato da OpenAI) è in grado di farlo con compiti legati al linguaggio, ma non c'è motivo per cui questo non debba essere possibile con immagini e suoni, o con combinazioni di questi tre elementi (Brown, 2020).

Inoltre, il paradosso di Moravec implica che lo sviluppo di "partner" di IA con molti tipi di qualità umane (o umanoidi) sarà molto difficile da ottenere, mentre il loro valore aggiunto (cioè al di là dei confini delle capacità umane) sarà relativamente basso. Le applicazioni più fruttuose dell'IA riguarderanno principalmente l'integrazione dei vincoli e delle limitazioni umane.

Dati gli attuali incentivi al progresso tecnologico competitivo, le forme multiple di sistemi di IA (connessi) ristretti saranno il principale motore dell'impatto dell'IA sulla nostra società nel breve e medio termine. Per il prossimo futuro, ciò potrebbe implicare che le applicazioni di IA rimarranno molto diverse e per molti aspetti quasi incomparabili con gli umani. Questo è probabile che sia vero anche se l'ipotetica corrispondenza dell'intelligenza artificiale generale (AGI) con la cognizione umana dovesse essere raggiunta più avanti nel tempo. L'intelligenza è un concetto multidimensionale (quantitativo e qualitativo). Tutte le dimensioni dell'intelligenza artificiale si sviluppano e crescono lungo un percorso diverso, con dinamiche proprie. Pertanto, nel tempo un numero crescente di capacità specifiche (ristrette) dell'IA potrebbe gradualmente eguagliare, superare e trascendere le capacità cognitive umane. Dati gli enormi vantaggi dell'IA, ad esempio quanto a disponibilità di dati e capacità di elaborazione dei dati, la realizzazione dell'IA probabilmente surclasserebbe l'intelligenza umana sotto molti aspetti. Ciò implica che l'ipotetico punto di incontro tra le capacità cognitive umane e quelle artificiali, cioè l'AGI di livello umano, sarà probabilmente difficile da definire (Goertzel, 2007). (3)

Quindi, quando l'IA ci concepirà veramente come un "amico", un "partner", un "alter ego" o un "compagno", come facciamo noi quando collaboriamo con altri esseri umani in quanto umani, ci supererà in molte aree allo stesso tempo (Moravec, 1998). Avrà un profilo di capacità e abilità completamente diverso e quindi non sarà facile capire davvero il modo in cui "pensa" e prende le sue decisioni. Nel frattempo, però, man mano che le capacità dei robot si espandono e passano da semplici strumenti a sistemi più integrati, è importante calibrare adeguatamente le nostre aspettative e percezioni nei confronti dei robot. Dovremo quindi migliorare la nostra consapevolezza e la nostra visione del continuo sviluppo e della progressione di molteplici forme di sistemi di intelligenza artificiale (integrati). Ciò riguarda, ad esempio, la natura sfaccettata dell'intelligenza. Diversi agenti possono avere combinazioni di intelligenze di livello molto diverso. Un

agente con intelligenza generale può, ad esempio, essere dotato di eccellenti capacità nell'area del riconoscimento delle immagini e della navigazione, del calcolo e del ragionamento logico e, allo stesso tempo, essere inadeguato nell'area dell'interazione sociale e della risoluzione di problemi orientati agli obiettivi. Questa consapevolezza della natura multidimensionale dell'intelligenza riguarda anche il modo in cui dobbiamo affrontare (*e* capitalizzare) l'antropomorfismo. Ovvero la tendenza umana nell'interazione uomo-robot a caratterizzare gli artefatti non umani che superficialmente ci assomigliano come se possedessero tratti, emozioni e intenzioni simili a quelli umani (ad esempio, Kiesler e Hinds, 2004; Fink, 2012; Haring et al., 2018). La comprensione di questi fattori umani è fondamentale per ottimizzare l'utilità, le prestazioni e la sicurezza dei sistemi umani-IA (Peeters et al., 2020).

Da questo punto di vista, la domanda se "l'intelligenza artificiale a livello umano" sarà realizzata o meno non è la questione più rilevante per il momento. Secondo la maggior parte degli scienziati che si occupano di IA, questo accadrà sicuramente e la domanda chiave non è se accadrà, ma QUANDO (vedi Müller e Bostrom, 2016). A livello di sistema, tuttavia, è probabile che molteplici applicazioni dell'IA ristretta superino l'intelligenza umana in sempre più settori.

5. CONCLUSIONI E QUADRO DI RIFERIMENTO

Il presente lavoro si è concentrato sul fornire una maggiore chiarezza e comprensione delle caratteristiche fondamentali, delle differenze e delle idiosincrasie delle intelligenze umane e artificiali. In primo luogo abbiamo presentato idee e argomenti per ampliare e differenziare la nostra concezione di intelligenza, sia essa umana o artificiale. Al centro di questa concezione più ampia e sfaccettata dell'intelligenza c'è l'idea che l'intelligenza in sé sia una questione di informazione e di calcolo, indipendente dal suo substrato fisico. Tuttavia, la natura di questo substra-

to fisico (biologico/carbonio o digitale/silicio) determinerà in modo sostanziale la dotazione di capacità cognitive e le relative limitazioni. Le facoltà cognitive organiche degli esseri umani si sono sviluppate molto recentemente nell'evoluzione della specie. Queste facoltà "embrionali" sono state costruite su un apparato di reti neurali biologiche ottimizzate per l'allostasi e le funzioni motorie percettive (complesse). La cognizione umana è quindi caratterizzata da varie limitazioni e distorsioni strutturali nella sua capacità di elaborare alcune forme di informazione non biologica. Le reti neurali biologiche, ad esempio, non sono molto capaci nell'eseguire calcoli aritmetici, per i quali la mia calcolatrice tascabile si adatta milioni di volte meglio. Queste limitazioni intrinseche e radicate, dovute all'origine biologica ed evolutiva dell'intelligenza umana, possono essere definite "radicate".

In linea con il *paradosso di Moravic*, abbiamo sostenuto che il comportamento intelligente è qualcosa di più rispetto a quello che *noi, come homo sapiens,* troviamo difficile. Non dobbiamo quindi confondere la difficoltà del compito (soggettiva, antropocentrica) con la complessità del compito (oggettiva). Abbiamo piuttosto sostenuto di concepire l'intelligenza in modo versatile e di riconoscerne le molteplici forme e possibili composizioni. Ciò implica un'elevata varietà di tipi di intelligenza biologica o di altre forme di intelligenza elevata (generale), con un'ampia gamma di possibili profili di intelligenza e qualità cognitive (che possono o meno superare la nostra in molti modi). Questo ci renderebbe più consapevoli delle potenzialità delle applicazioni dell'IA nel futuro a breve e medio termine. Per esempio, da questo punto di vista, la nostra ricerca dovrebbe concentrarsi principalmente su quelle componenti dello spettro dell'intelligenza che sono relativamente difficili per il cervello umano e relativamente facili per le macchine. Si tratta principalmente della componente *cognitiva* che richiede calcolo, analisi aritmetica, statistica, calcolo delle probabilità, analisi dei dati, ragionamento logico, memorizzazione, ecc.

In linea con questo abbiamo sostenuto una visione più umile, della nostra intelligenza umana generale. Ciò implica anche

che l'AGI di livello umano non dovrebbe essere considerata come l'ideale ultimo dell'intelligenza (da perseguire con la massima priorità). A causa delle molte differenze fondamentali tra intelligenze naturali e artificiali, un'intelligenza artificiale simile a quella umana sarà molto difficile da realizzare (e anche con un valore aggiunto relativamente limitato). Nel caso in cui un'AGI venga realizzata in un (lontano) futuro, essa avrà probabilmente un profilo di capacità cognitive e di abilità completamente diverso da quello di noi esseri umani. Quando una tale AGI sarà arrivata al punto di essere in grado di "collaborare" come un essere umano, è probabile che per molti aspetti sia già in grado di funzionare a livelli altamente superiori rispetto a quelli di cui siamo capaci noi. Per il momento, tuttavia, non sarà molto realistico e utile puntare a un'IA che includa l'ampia gamma di capacità percettive-motorie e cognitive umane. Invece, le applicazioni di IA più redditizie per il futuro a breve e medio termine saranno probabilmente basate su sistemi di IA multipli e ristretti. Queste applicazioni di IA multiple e ristrette potrebbero raggiungere l'intelligenza umana in una gamma sempre più ampia di settori.

Da questo punto di vista suggeriamo di non soffermarci troppo riflettendo se o quando l'IA ci supererà in astuzia, prenderà il nostro posto di lavoro o pensando a come dotarla di tutte le abilità umane. Considerato lo stato dell'arte, potrebbe essere saggio concentrarsi maggiormente sull'intero sistema di innovazioni multiple dell'IA con l'uomo come fattore chiave di collegamento e supervisione. Ciò implica anche la definizione e la formalizzazione di confini legali e di obiettivi adeguati (efficaci, etici e sicuri) per i sistemi di IA (Elands et al., 2019; Aliman, 2020). Quindi questo fattore umano (legislatore, utente, "collaboratore") deve avere una buona conoscenza delle caratteristiche e delle capacità dell'intelligenza biologica e artificiale (in tutti i tipi di compiti e condizioni di lavoro). Sia sul posto di lavoro che nella definizione delle politiche, le applicazioni dell'IA più proficue saranno quelle che integrano e compensano le limitazioni biologiche e cognitive intrinseche degli esseri umani. Per questo motivo, le questioni più importanti riguardano come usarla in modo intel-

ligente. Per quali compiti e in quali condizioni è sicuro lasciare le decisioni all'IA e quando è necessario il giudizio umano? Come sfruttare i punti di forza dell'intelligenza umana e come impiegare efficacemente i sistemi di IA per integrare e compensare i limiti intrinseci della cognizione umana? Per una panoramica recente si vedano (Hoffman e Johnson, 2019; Shneiderman, 2020a; Shneiderman, 2020b).

In sintesi: per quanto gli agenti di IA autonomi possano diventare intelligenti sotto certi aspetti, almeno nel prossimo futuro rimarranno macchine inconsapevoli. Queste macchine hanno un sistema operativo diverso (noi biologico, loro digitale) con capacità e qualità cognitive conseguentemente diverse rispetto alle persone e agli altri animali. Quindi, prima di avviare un "lavoro di squadra", i membri del team umano dovranno comprendere questo tipo di differenze, ossia come l'elaborazione delle informazioni e l'intelligenza umana differiscono da quelle delle numerose varianti possibili e specifiche dei sistemi IA. Solo quando gli esseri umani svilupperanno un'adeguata conoscenza di queste differenze "interspecie", potranno capitalizzare efficacemente i potenziali vantaggi dell'IA nei (futuri) team umani-IA. Data l'elevata flessibilità, versatilità e adattabilità degli esseri umani rispetto ai sistemi di IA, la prima sfida diventa quindi: come garantire l'adattamento umano alle capacità più rigide dell'IA? (4) In altre parole: come possiamo concepire correttamente le differenze tra intelligenza umana e artificiale?

5.1 QUADRO DI RIFERIMENTO PER LA FORMAZIONE SULLA CONSAPEVOLEZZA DELL'INTELLIGENZA

Per questo motivo, la questione della *consapevolezza dell'intelligenza* nei professionisti umani deve essere affrontata meglio. Oltre agli strumenti informatici per diffondere informazioni rilevanti sulla consapevolezza (Collazos et al., 2019) nei sistemi uomo-macchina, ciò richiede una migliore istruzione e formazione su come affrontare le caratteristiche, le idiosincrasie e le capacità nuove e diverse dei sistemi di IA. Ciò include, ad esempio, un'adeguata comprensione delle caratteristiche di base, del-

le possibilità e dei limiti delle proprietà del sistema cognitivo dell'IA, senza idee sbagliate antropocentriche e/o antropomorfiche. In generale, questa "*consapevolezza dell'intelligenza*" è molto importante per comprendere, studiare e affrontare meglio le molte possibilità e sfide dell'intelligenza artificiale. Questa sfida pratica dei fattori umani potrebbe essere affrontata, ad esempio, sviluppando percorsi di formazione e ambienti di apprendimento nuovi, mirati e facilmente configurabili (adattivi) per i sistemi umano-IA. Queste forme e ambienti di formazione flessibili (ad esempio, simulazioni e giochi) dovrebbero concentrarsi sullo sviluppo di conoscenze, intuizioni e abilità pratiche riguardanti le caratteristiche, le abilità e i limiti specifici e non umani dei sistemi di intelligenza artificiale e su come affrontarli in situazioni pratiche. Le persone dovranno comprendere i fattori critici che determinano gli obiettivi, le prestazioni e le scelte dell'IA? In alcuni casi questi possono anche comprendere la semplice nozione che le IA si entusiasmano per le loro prestazioni nel raggiungere i loro obiettivi come il vostro frigorifero si entusiasma per la conservazione del vostro frullato. Devono imparare quando e in quali condizioni è sicuro delegare le decisioni all'IA e quando invece è necessario il giudizio umano. E più in generale: come "pensa" e decide? L'importanza di questo tipo di conoscenze, competenze e pratiche diventerà maggiore solo quando crescerà il grado di autonomia (e generalismo) dei sistemi di IA avanzati.

Come si presenta un programma di formazione *sulla consapevolezza dell'intelligenza*? Deve includere almeno un modulo sulle caratteristiche cognitive dell'IA. Si tratta di un argomento simile a quelli che sono inclusi nei programmi di studio sulla cognizione *umana*. Questo ampio modulo sulla "Scienza cognitiva dell'IA" può comprendere una serie di sotto-argomenti, a partire da una revisione del concetto di "Intelligenza", emendato dagli equivoci antropocentrici e antropomorfi. Inoltre, questo modulo dovrebbe concentrarsi sulla conoscenza della struttura e del funzionamento del sistema operativo dell'IA o della "mente dell'IA". A questo possono seguire argomenti come: percezione e interpretazione delle informazioni da parte dell'IA, cognizione

dell'IA (memoria, elaborazione delle informazioni, risoluzione dei problemi, pregiudizi), gestione delle possibilità e dei limiti dell'IA nelle aree "umane" come la creatività, l'adattabilità, l'autonomia, la riflessione e la (auto)consapevolezza, gestione delle funzioni obiettivo (valutazione delle azioni in relazione ai costi-benefici), etica dell'IA e sicurezza dell'IA. Inoltre, tale curriculum dovrebbe comprendere dei moduli tecnici che forniscano una visione del funzionamento del sistema operativo dell'IA. Data l'enorme velocità con cui vengono sviluppate la tecnologia e le applicazioni dell'IA, anche il contenuto di tale curriculum è molto dinamico e in continua evoluzione in linea con il progresso tecnologico. Ciò implica che il curriculum e gli ausili e gli ambienti di formazione devono essere flessibili, esperienziali e adattivi, il che rende la forma di lavoro del *serious gaming* particolarmente adatta. Di seguito, forniamo un quadro globale per lo sviluppo di nuovi curriculum educativi sulla consapevolezza dell'IA. Questi argomenti vanno al di là dell'imparare a "operare", "controllare" o interagire con applicazioni specifiche dell'IA (cioè l'interazione uomo-macchina convenzionale):

- Comprensione delle caratteristiche del sistema sottostante l'IA (il "cervello dell'IA"). Comprendere le qualità e i limiti specifici dell'IA rispetto all'intelligenza umana.
- Comprendere la complessità dei compiti e dell'ambiente dal punto di vista dei sistemi di intelligenza artificiale.
- Comprendere il problema dei pregiudizi nella cognizione umana, rispetto ai pregiudizi nell'IA.
- Comprendere i problemi associati al controllo dell'IA, alla prevedibilità del comportamento dell'IA (decisioni), alla costruzione della fiducia, al mantenimento della consapevolezza della situazione (compiacenza), all'assegnazione dinamica dei compiti (ad esempio, l'assunzione dei compiti dell'altro) e alla responsabilità (accountability).
- Come affrontare le possibilità e i limiti dell'IA nel campo della "creatività", dell'adattabilità dell'IA, della "consapevolezza ambientale" e della generalizzazione delle nozioni.

- Imparare a gestire i limiti percettivi e cognitivi e i possibili errori dell'IA che possono essere difficili da comprendere.
- Fiducia nelle prestazioni dell'IA (eventualmente nonostante la limitata trasparenza o capacità di "spiegare") basata su verifica e validazione.
- Imparare a gestire la nostra naturale inclinazione all'antropocentrismo e all'antropomorfismo ("teoria della mente") quando ragioniamo sull'interazione uomo-robot.
- Capire come capitalizzare i poteri dell'IA per affrontare i limiti intrinseci dell'elaborazione delle informazioni da parte dell'uomo (e viceversa).
- Comprendere le caratteristiche e le qualità specifiche del sistema uomo-macchina ed essere in grado di decidere quando, per cosa e come la combinazione integrata di facoltà umane e IA può funzionare al meglio delle potenzialità complessive del sistema.

In conclusione: a causa dell'enorme velocità con cui si evolvono la tecnologia e le applicazioni dell'IA, abbiamo bisogno di una concettualizzazione più versatile dell'intelligenza e di un riconoscimento delle sue molte forme e combinazioni possibili. Una concezione riveduta dell'intelligenza comprende anche una buona comprensione delle caratteristiche di base, delle possibilità e dei limiti delle proprietà dei diversi sistemi cognitivi (biologici e artificiali), senza fraintendimenti antropocentrici e/o antropomorfi. Questa "consapevolezza dell'intelligenza" è molto importante per comprendere meglio e affrontare le molteplici possibilità e sfide dell'intelligenza artificiale, ad esempio per decidere quando utilizzare o impiegare l'IA in relazione ai compiti e al loro contesto. Si raccomanda pertanto lo sviluppo di programmi educativi con forme di formazione e ambienti di apprendimento nuovi, mirati e facilmente configurabili per i sistemi umani-IA. Ulteriori lavori dovrebbero concentrarsi su strumenti, metodi e contenuti formativi sufficientemente flessibili e adattivi da poter tenere il passo con i rapidi cambiamenti nel campo dell'IA e con l'ampia varietà di gruppi target e obiettivi di apprendimento.

2 - ESTRAZIONE DI DATI MEDICI BASATA SULL'INTELLIGENZA ARTIFICIALE

Tratto e tradotto da
Amjad Zia, Muzzamil Aziz, Ioana Popa, Sabih Ahmed Khan, Amirreza Fazely Hamedani e Abdul R. Asif, *Artificial Intelligence-Based Medical Data Mining*, J. Pers. Med. 2022, 12(9), 1359.

https://doi.org/10.3390/jpm12091359

Le parti omesse dal curatore rispetto all'originale sono indicate dal segno **[...]**

ABSTRACT

La comprensione dei dati testuali non strutturati pubblicati utilizzando approcci e strumenti tradizionali di text mining sta diventando una sfida a causa del rapido aumento delle pubblicazioni elettroniche open-source. L'applicazione di tecniche di data mining nelle scienze mediche è una tendenza emergente; tuttavia, gli approcci tradizionali di text mining [letteralmente: "estrazione di testi"] sono insufficienti per far fronte all'attuale aumento del volume di dati pubblicati. Pertanto, vengono sviluppati e utilizzati strumenti di text mining basati sull'intelligenza artificiale per elaborare grandi volumi di dati ed esplorare le caratteristiche e le correlazioni nascoste nei dati. Questa rassegna fornisce una comprensione chiara e approfondita di come la tecnologia di data mining basata sull'intelligenza artificiale viene utilizzata per analizzare i dati medici. Viene inoltre descritto un processo standard di data mining basato sul CRISP-DM (Cross-Industry Standard Process for Data Mining) e gli strumenti/librerie più comuni disponibili per ogni fase del data mining medico.

1. INTRODUZIONE

Con la rapida crescita della letteratura medica disponibile online, è difficile per i lettori ottenere le informazioni desiderate senza un grande investimento di tempo. Ad esempio, nella pandemia COVID-19 in corso, il numero di pubblicazioni che parlano di COVID-19 è aumentato molto rapidamente. Nei primi due anni della pandemia, ci sono stati 228.640 articoli in PubMed, 282.883 articoli in PMC e 7551 studi clinici COVID-19 elencati nei database ClinicalTrials.gov (dati consultati il 16 febbraio 2022), e la loro velocità di crescita è sorprendente. A causa dell'elevato grado di eterogeneità dimensionale, irregolarità e tempestività, questi dati sono spesso sottoutilizzati. Questa crescita esponenziale della letteratura scientifica ha reso difficile per i ricercatori (i) ottenere informazioni rilevanti dalla letteratura, (ii) presentare le informazioni in modo conciso e strutturato da un ammasso di letteratura non strutturata e (iii) comprendere appieno lo stato attuale e la direzione dello sviluppo in un campo di ricerca.

La letteratura in rapido aumento non può essere gestita e/o elaborata con tecnologie e metodi tradizionali in tempi accettabili. Questo enorme volume di dati rende piuttosto difficile per i ricercatori esplorare, analizzare, visualizzare e ottenere un risultato conciso. Il processo di estrazione di modelli nascosti, significativi e vincolanti dalla letteratura testuale non strutturata è noto come text mining [1]. Le tecniche tradizionali di text mining non sono sufficienti a gestire gli attuali grandi volumi di letteratura pubblicata. Pertanto, si intravede all'orizzonte un rapido aumento dello sviluppo di nuove tecniche di data mining basate sull'intelligenza artificiale, a beneficio di pazienti e medici. Includere l'intelligenza artificiale (anche l'apprendimento automatico (ML [*Machine Learning*]), l'apprendimento profondo (DL [*Deep Learning*]) e l'elaborazione del linguaggio naturale (NLP [*Natural Language Processing*]) come sottoinsiemi) potenzia il processo di data mining con vari vantaggi: l'acquisizione di nuove conoscenze nel processo decisionale, l'elaborazione di grandi serie di dati

con maggiore accuratezza ed efficienza e la capacità di apprendere e migliorare continuamente dai nuovi dati.

La presente rassegna fa luce sul ruolo dei diversi metodi basati sull'IA, ossia NLP e reti neurali (NN [*Neural Networks*]) nell'estrazione di testi medici, sugli attuali processi di data mining, sulle diverse fonti di database e sui vari strumenti basati sull'IA utilizzati nel processo di text mining insieme a diversi algoritmi. Abbiamo esaminato i più recenti approcci al text mining, evidenziato le differenze chiave tra il data mining medico e non medico e presentato una serie di strumenti e tecniche attualmente utilizzati per ogni fase del text mining della letteratura medica. Inoltre, abbiamo descritto il ruolo dell'intelligenza artificiale e dell'apprendimento automatico nell'estrazione dei dati medici e abbiamo evidenziato le sfide, le difficoltà e le opportunità rilevate lungo il percorso.

1.1 TEXT MINING DI LETTERATURA MEDICA E NON MEDICA

I dati medici umani sono unici e possono essere difficili da estrarre e analizzare. In primo luogo, dato che gli esseri umani sono la specie più avanzata e più osservata (e meglio) del pianeta, la loro osservazione è arricchita perché gli esseri umani possono fornire facilmente i loro input sensoriali rispetto alle altre specie sulla terra [2]. Tuttavia, l'estrazione di dati medici si trova ad affrontare numerose sfide fondamentali, principalmente a causa dell'eterogeneità e della verbosità dei dati provenienti da varie cartelle cliniche non standardizzate. Anche la scarsa qualità dei dati è un problema noto nella scienza medica, che deve essere gestito con cura per l'estrazione dei dati. Queste sfide possono essere affrontate standardizzando il processo di selezione dei pazienti, la raccolta, l'archiviazione, l'annotazione e la gestione dei dati [3]. Tuttavia, a volte ciò significa che i dati esistenti e quelli acquisiti in più centri senza un buon coordinamento e procedure operative standard (SOP) non possono essere utilizzati. La principale divergenza tra l'estrazione di dati medici e non medici riguarda gli aspetti etici e legali. L'uso di informazioni che pos-

sono essere ricondotte a singoli individui comporta rischi per la privacy, con conseguenti problemi legali. Più di quindici dipartimenti federali statunitensi con il Dipartimento della Salute e dei Servizi Umani degli Stati Uniti hanno pubblicato le revisioni finali della politica federale per la protezione dei soggetti umani "la Common Rule, 45 CFR 46, Subpart A" (Protection of Human Subjects, 45 CFR 46 (2018). Il quadro federale per la privacy e la sicurezza non si applica alle informazioni de-identificate o anonimizzate [4].

La proprietà dei dati medici è un'altra questione critica, poiché i dati vengono acquisiti da diverse entità in cui gli individui possono essere stati durante il trattamento o per scopi diagnostici. Questi enti possono raccogliere e conservare i dati in base all'autorizzazione dell'individuo al momento dell'acquisizione dei dati. Tuttavia, tale autorizzazione può essere ritirata dal paziente in qualsiasi momento e/o il consenso è valido solo per un periodo limitato e i dati devono essere cancellati dopo tale periodo [5]. La maggior parte dei testi clinici è scritta con poche informazioni utili ma anche tante informazioni extra. Inoltre, sono scritti per il personale clinico e i colleghi, quindi sono pieni di frasi incomplete e abbreviazioni. Per leggere, comprendere ed elaborare questi testi sono necessari strumenti speciali [6]. Le cartelle cliniche elettroniche, note anche come testi clinici, presentano un solo problema: sono scritte in un linguaggio altamente specializzato che può essere elaborato con pochi strumenti attualmente disponibili. In secondo luogo, le cartelle cliniche sono talvolta scritte in uno stile telegrafico e denso di informazioni per la comunicazione tra medici, e non esiste un dizionario sviluppato per tali comunicazioni per controllare gli errori grammaticali e ortografici. Inoltre, i medici e il personale sanitario usano spesso frasi incomplete e spesso non menzionano il soggetto, ad esempio il paziente, perché è sottinteso. "Arrivato con 38,3 di febbre e polso a 132", ad esempio.

1.2 USO DELL'INTELLIGENZA ARTIFICIALE E DELL'APPRENDIMENTO AUTOMATICO NELL'ESTRAZIONE DEI DATI DELLA LETTERATURA MEDICA

L'era digitale ha dimostrato una fiducia immensa e crescente nelle tecniche di apprendimento automatico [*Machine Learning*, ML] per aumentare la qualità della vita in quasi tutti i campi della vita. È il caso dell'assistenza sanitaria e della medicina di precisione, dove un flusso continuo di dati medici provenienti da fonti eterogenee diventa un fattore chiave per i trattamenti e le diagnosi assistite dall'IA/ML. Ad esempio, oggi l'IA può aiutare i medici a migliorare i risultati dei pazienti con diagnosi e piani di trattamento precoci e a migliorare la qualità della vita. Allo stesso modo, le organizzazioni e le autorità sanitarie vorrebbero poter usare presto l'IA per la prognosi di epidemie e pandemie a livello nazionale e internazionale. Anche nel settore sanitario si assiste oggi all'utilizzo di procedure assistite dall'intelligenza artificiale per la gestione operativa, sotto forma di documentazione automatizzata, programmazione di appuntamenti e assistenza virtuale per i pazienti. In questa sezione vedremo alcuni esempi concreti di strumenti e tecnologie AIML attualmente utilizzati in varie aree delle scienze mediche (Tabella 1).

Prodotti/prototipi di ricerca	Trattamento/Campo di studio	Azienda/Istituzione	Riferimento
UnisciPACS™	Radiologia clinica per immagini	IBM Watson	Panoramica del PACS Merge | IBM
BiometryAssist	Ecografia diagnostica	Samsung Medison	https://www.intel.com/content/www/us/en/developer/tools/oneapi/application-catalog/full-catalog/diagnostic-ultrasound.html (consultato il 17 febbraio 2022)
LaborAssist™	Ecografia diagnostica	Samsung Medison	

Soluzione per il rilevamento del cancro al seno	Ecografia, mammografia, risonanza magnetica	La soluzione di Huiying	https://builders.intel.com/ai/solutionscatalog/breast-cancer-detection-solution-657 (consultato il 17 febbraio 2022)
Soluzione TC	Rilevazione precoce di COVID-19	La soluzione di Huiying	https://builders.intel.com/ai/solutionscatalog/ct-solution-for-early-detection-of-covid-19-704 (consultato il 17 febbraio 2022)
Sistema CAD per polmonite TAC Dr. Pecker	Classificazione e quantificazione di COVID-19	Tecnologia Jianpei	https://www.intel.com/content/www/us/en/developer/tools/oneapi/application-catalog/full-catalog/dr-pecker-ct-pneumonia-cad-system.html (consultato il 17 febbraio 2022)

Tabella 1. Prodotti e prototipi di ricerca AIML di alcune organizzazioni leader nel settore sanitario.

Prima di entrare nel dettaglio, vale la pena ricordare che i concetti di data mining e machine learning vanno di pari passo e si sovrappongono in una certa misura, ma con una chiara distinzione nel risultato finale. Il data mining è il processo di scoperta di correlazioni, anomalie e nuovi modelli in un ampio insieme di dati provenienti da un esperimento o da un evento per prevedere i risultati [7]. La base del data mining è costituita da tecniche di modellazione statistica che rappresentano i dati in un modello matematico ben definito e poi usano questo modello per creare relazioni e altri modelli tra le variabili dei dati. L'apprendimento automatico, d'altra parte, è un approccio al data mining che consente agli algoritmi di apprendimento automatico di far comprendere i dati al computer (con l'aiuto di modelli statistici) e di fare previsioni proprie. Detto ciò, le tecniche di data mining richiedono sempre l'interazione umana per trovare

modelli di interesse da un set di dati, mentre l'apprendimento automatico è una tecnica relativamente moderna che consente ai programmi informatici di imparare dai dati automaticamente e di fornire previsioni senza alcuna interazione umana.

1.3 ELABORAZIONE DEL LINGUAGGIO NATURALE

L'elaborazione del linguaggio naturale (NLP) è una disciplina dell'intelligenza artificiale (AI) che converte il linguaggio umano in linguaggio macchina. Con la crescita nell'uso della tecnologia informatica negli ultimi 20 anni, questo settore è cresciuto in modo significativo [8]. La documentazione clinica, il riconoscimento vocale, la codifica assistita da computer, il data mining, la reportistica automatizzata dei registri, il supporto alle decisioni cliniche, l'associazione di studi clinici, l'autorizzazione preventiva, i chatbot AI e gli strumenti di scrittura virtuale, i modelli di aggiustamento del rischio, la fenotipizzazione computazionale, la gestione delle revisioni e l'analisi del sentiment, la dettatura e le implementazioni EMR e l'analisi delle cause profonde sono alcune delle applicazioni più diffuse dell'NLP in ambito sanitario [9]. In letteratura è stata illustrata un'ampia gamma di applicazioni della PNL.

Liu et al. [10] hanno utilizzato un testo clinico per il riconoscimento di entità utilizzando tecniche di word embedding (WE)-skipgram e memoria a breve termine (LSTM) e hanno ottenuto un'accuratezza del 94,37%, del 92,29% e dell'85,81% rispettivamente per la de-identificazione, il rilevamento di eventi e l'estrazione di concetti, in base al punteggio F1 medio. Deng et al. [11] hanno utilizzato il concept embedding (CE)-continuous bag of words (CBOW), lo skip-gram e la proiezione casuale per generare rappresentazioni di codice e semantiche da testi clinici. Afzal et al. [12] hanno sviluppato una pipeline per la generazione di domande, il riconoscimento della qualità delle prove, la classificazione e la sintesi delle prove dalla letteratura biomedica e hanno ottenuto un'accuratezza del 90,97%. Oltre a questi esempi, Pandey et al. [13] hanno elencato 57 lavori pubblicati tra il 2017 e il 2019 che hanno utilizzato tecniche NLP e varie

fonti testuali, come testi clinici, input EHR, testi medici cinesi, rapporti di patologia oncologica, testi biomedici, articoli di studi controllati randomizzati (RCT), note cliniche e rapporti radiologici di testi EMR, tra gli altri.

2. PROCESSO STANDARD PER L'ESTRAZIONE DEI DATI

In risposta alla richiesta di un metodo standard di data mining, i leader del settore hanno collaborato con un gruppo eterogeneo di professionisti (fornitori di servizi, consulenti di gestione, utenti di data mining, venditori di data warehouse) ed esperti di data mining per sviluppare un modello di data mining gratuito, ben documentato e non proprietario [14]. Per il data mining sono disponibili numerosi metodi, come ASUM (*Analytics Solutions Unified Method*), CRISP-DM (*Cross-Industry Standard Process for Data Mining*), KDD (*Knowledge discovery in databases*), SEMMA (*Sampling, Exploring, Modifying, Modelling, and Assessing*) e POST-DS (*Process Organization and Scheduling electing Tools for Data Science*) [15]. In questo studio, utilizziamo il modello CRISP-DM per il data mining perché è un approccio completo ed esaustivo. Nel 1997, la partnership CRISP-DM ha sviluppato un modello generico di processo per l'estrazione dei dati per stabilire linee guida per i principianti del data mining, la comunità e gli esperti, che possono essere modificate per qualsiasi esigenza particolare [14]. Ad esempio, per affrontare il problema delle serie temporali multidimensionali in un'unità di terapia intensiva neonatale (NICU), il modello CRISP-DM è stato modificato per supportare e accogliere il data mining temporale (TDM), che è stato denominato CRISP-TDM [16]. Nel ciclo di vita di un processo di data mining, il modello di riferimento CRISP-DM prevede sei fasi (Figura 1): comprensione del business, comprensione dei dati, preparazione dei dati, modellazione, valutazione e implementazione. I dettagli degli strumenti e delle tecnologie disponibili per ciascuna fase sono descritti nel resto dell'articolo.

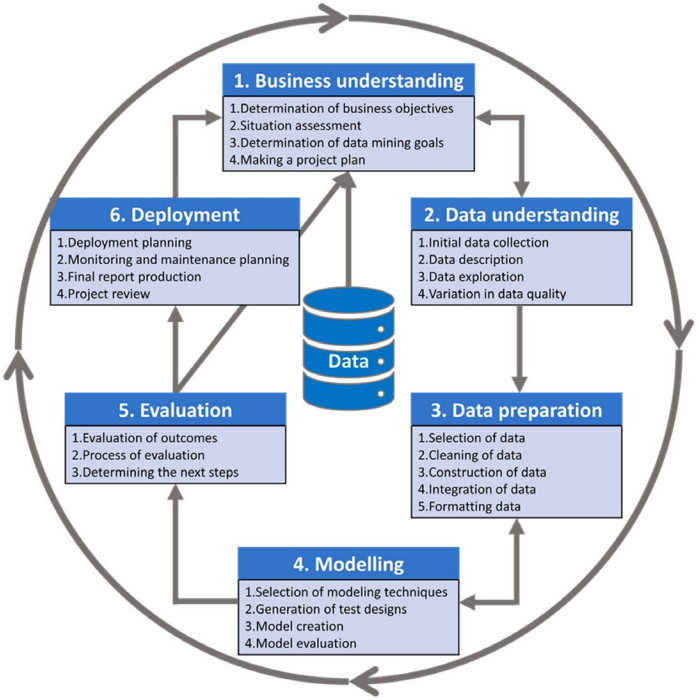


Figura 1. Processo standard intersettoriale per l'estrazione dei dati (CRISP-DM) – adattato dalla pagina web della Data Science Process Alliance [17] (www.datascience-pm.com/crisp-dm-2/, consultato il 16 aprile 2022). La natura circolare del processo di data mining è indicata dal cerchio esterno, mentre le frecce che collegano le fasi mostrano le relazioni più importanti e comuni.

2.1 COMPRENSIONE DEL BUSINESS

La prima e più importante parte del data mining è la comprensione del business, che comprende la definizione degli obiettivi e dei target del progetto, la valutazione della situazione, i piani di esecuzione e la valutazione dei rischi [14]. La definizione degli obiettivi del progetto richiede una comprensione completa del vero scopo del progetto per definire le variabili associate. Le fasi della comprensione dei dati secondo il CRISP-DM sono: (1) determinare gli obiettivi aziendali (per comprendere appieno l'obiettivo del progetto, identificare gli attori chiave e stabilire i

criteri di successo aziendali), (2) valutare la situazione (per identificare la disponibilità di risorse (in particolare i dati), identificare i rischi del progetto e le potenziali soluzioni a tali rischi e calcolare il rapporto costi-benefici), (3) chiarire gli obiettivi del data mining (per stabilire gli obiettivi del progetto e i criteri di successo), (4) produrre un piano di progetto (per sviluppare piani dettagliati per ogni segmento del progetto, tra cui una tempistica e la selezione di tecnologie e strumenti).

Martins et al. [18] hanno utilizzato un approccio di data mining per prevedere le malattie cardiovascolari (utilizzando i software RapidMiner e Weka). La questione principale affrontata dal progetto è come individuare precocemente le malattie cardiovascolari in una persona ad alto rischio di sviluppare la malattia ed evitare così una morte prematura. Di conseguenza, l'obiettivo principale è quello di creare una soluzione per la previsione delle malattie cardiovascolari nei pazienti utilizzando i dati dei pazienti, per ridurre il tempo necessario per la diagnosi della malattia e per fornire ai pazienti un trattamento immediato e adeguato.

2.2 COMPRENSIONE DEI DATI

L'enfasi di questa fase (seconda fase), secondo il CRISP-DM, è sull'identificazione della fonte dei dati, sull'acquisizione dei dati, sulla raccolta iniziale dei dati, sulla familiarizzazione con i dati e sull'identificazione dei problemi nei dati acquisiti. Le fasi della fase di comprensione dei dati sono: (1) acquisire i dati iniziali (raccogliere i dati da varie fonti, inserirli nel programma di analisi e integrarli), (2) spiegare i dati (studiare e riferire sulle proprietà superficiali dei dati acquisiti, come l'identità del campo, il formato dei dati, la quantità di dati e il numero di record, ecc.), (3) esplorare i dati (per approfondire i dati interrogandoli, visualizzandoli e identificando le relazioni tra i punti di dati, nonché per generare un rapporto di esplorazione) e (4) verificare la qualità dei dati (per ispezionare e documentare la qualità dei dati ed eventuali problemi legati alla qualità) [14]. In questa fase ci si concentra sull'identificazione delle fonti di dati per i vari tipi di dati, sul processo di acquisizione dei dati e sulla gestione

delle restrizioni di accesso nell'acquisizione dei dati. L'industria sanitaria e le istituzioni mediche generano ogni giorno un'enorme quantità di dati provenienti dalla diagnostica per immagini, dal monitoraggio dei pazienti e dalle cartelle cliniche [7]. Alcuni dei tipi più comuni di dati medici sono i dati sperimentali, la letteratura medica, i dati clinici testuali, le cartelle cliniche, le immagini/video (ad esempio, la risonanza magnetica) e i dati omici (ad esempio, genomici e proteomici). Ad esempio, Martins et al. [18] hanno utilizzato un approccio di data mining per prevedere le malattie cardiovascolari. Per la comprensione dei dati, il set di dati per la previsione delle malattie cardiovascolari proveniva dal repository di dati Kaggle e si concentrava sulla rilevazione di casi di malattie cardiovascolari. Il set di dati comprendeva 70.000 pazienti registrati con 12 attributi relativi alla malattia raccolti durante le visite mediche dei pazienti.

2.2.1 Estrazione della letteratura/raccolta di dati

Il primo compito della fase di comprensione dei dati è identificare le fonti di dati, acquisire i dati da queste fonti, identificare i problemi durante l'acquisizione dei dati, come le restrizioni sui dati e le politiche sulla privacy, e documentare le soluzioni [14]. Il text/data mining utilizza spesso fonti pubbliche su Internet, come il World Wide Web. Il recupero di contenuti da fonti pubbliche viene definito "web scraping" o "web crawling". Il web scraping può essere eseguito manualmente, ma anche automaticamente con l'aiuto di un web crawler. Lo scraping manuale di un grande database come PubMed, che contiene milioni di pubblicazioni peer-reviewed, richiede molto tempo e fatica. Solo l'elaborazione automatizzata può fornire la qualità, il tempo di risposta e l'omogeneità necessari per l'analisi di un database così ampio. Di conseguenza, c'è sempre una forte richiesta di tecniche e strumenti di web scraping adattati alle esigenze dei clienti. PubMed, ad esempio, è un enorme database di letteratura biomedica che contiene 34 milioni di citazioni (all'11 maggio 2022) raccolte da libri online, riviste di scienze della vita e MEDLINE, e ogni anno

viene aggiunto un numero enorme di nuove pubblicazioni [19]. I web crawler sono utilizzati per cercare e raccogliere i dati necessari. Guo et al. [20], ad esempio, hanno raccolto i dati COVID-19 pubblicati dalle autorità sanitarie locali utilizzando un web crawler (sviluppato con il linguaggio Python e collegato a un database MySQL).

Sebbene il web scraping e il web crawling possano sembrare identici, presentano diverse distinzioni (Figura 2). Infatti i termini "web scraping" e "web crawling" vengono talvolta scambiati, ma si riferiscono a due processi distinti [21,22]. Web crawling è un termine ampio che indica il processo di scaricamento di informazioni da un sito web, l'estrazione dei collegamenti ipertestuali in esso contenuti e il loro seguito (Figura 2). In genere, le informazioni scaricate vengono salvate in un database o indicizzate per consentire la ricerca. In sostanza, i motori di ricerca sono dei crawler. Tutto ciò che serve è vedere una pagina nella sua interezza e indicizzarla. Quando un bot effettua il crawling di un sito web, analizza ogni pagina e ogni link, fino all'ultima riga del sito, alla ricerca di qualsiasi informazione. I web crawler sono utilizzati per lo più dai principali motori di ricerca come Google, Bing e Yahoo, oltre che da istituzioni di statistica e aggregatori online. In genere, un web crawler raccoglie informazioni generali, mentre gli scrapers raccolgono insiemi di dati particolari [23,24]. D'altra parte, il web scraping è il processo di ottenimento di dati da una pagina web e di estrazione di informazioni specifiche che possono essere salvate praticamente ovunque (database, file, ecc.), come illustrato nella Figura 2. Uno scraper online, noto anche come estrattore di dati web, è simile a un web crawler in quanto rileva e individua i contenuti di un sito web. A differenza di un web crawler, che utilizza ID pseudocasuali, il web scraping utilizza identificatori specifici, come la struttura HTML delle pagine web da cui devono essere raccolti i dati. Il web scraping si riferisce all'uso di robot per estrarre da Internet insiemi di dati specifici. I dati ottenuti possono essere confrontati, controllati e analizzati in base alle esigenze e agli obiettivi dell'organizzazione che li sta cercando [25].

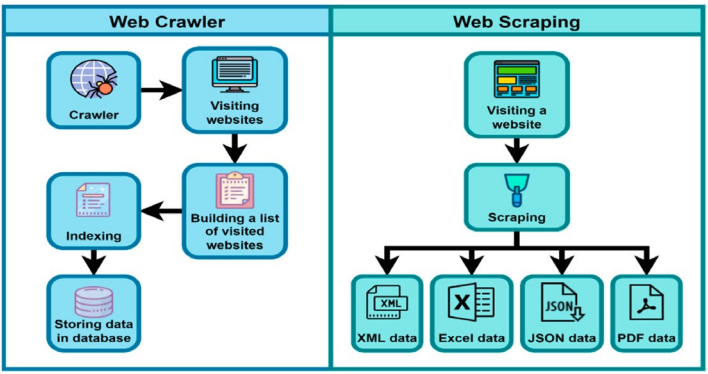


Figura 2. Confronto tra web crawling e web scraping.

Oggi sono disponibili numerosi strumenti di text mining. Kaur e Chopra [26] hanno confrontato 55 strumenti di text mining popolari e le loro caratteristiche, dividendoli in tre categorie: (1) strumenti proprietari (di proprietà dell'azienda), (2) strumenti open source (gratuiti) e (3) strumenti di text mining online (eseguiti direttamente da un sito web). Nella Tabella 2 sono riportati quattro strumenti che non erano stati esaminati nella precedente revisione, ma che ora fanno parte dell'elenco degli strumenti di text mining più apprezzati. Tutti questi strumenti basati su Python servono allo stesso scopo, ma con finalità e obiettivi diversi. Il vantaggio di Requests rispetto agli altri strumenti è la sua facilità d'uso, che lo rende una scelta eccellente per qualsiasi attività semplice di scraping del web. Scrapy è più adatto per progetti di web scraping su larga scala, al contrario degli altri tre strumenti (requests, beautiful soup e selenium), che sono più adatti per attività di scraping su piccola scala. Lo strumento Beautiful Soup presenta i vantaggi della semplicità di comprensione, apprendimento e utilizzo e della possibilità di estrarre informazioni da un sito web disorganizzato. Selenium ha un vantaggio significativo rispetto agli altri strumenti di scraping descritti, perché è in grado di eseguire lo scraping di siti web con un codice JavaScript pesante. La Tabella 2 fornisce le descrizioni dei confronti gerarchici.

	Requests	Scrapy	Beautiful Soup	Selenium
Che cos'è?	Libreria HTTP per Python	Framework web open-source scritto in Python	libreria python	Tool framework per applicazioni open-source e libreria python
Obiettivo	Invio di richieste HTTP/1.1 con Python	Può effettuare il crawling o lo scraping di siti web ed estrarre i dati strutturati e salvarli. Può essere utilizzato anche per un'ampia gamma di attività, monitoraggio e test automatizzati.	Può analizzare i dati e fare lo scraping delle pagine web. Estrarre informazioni da documenti XML e HTML	Utile per lo scraping di siti web che sono pesanti per via di JavaScript
Utilizzo ideale	Utilizzato per attività di scraping web semplici e complesse di basso livello	Framework utilizzato per attività complesse di web scraping o web crawling. Utilizzato per progetti su larga scala	Utilizzato per piccole attività di scraping del web Toolkit per la ricerca in un documento (XML o HTML) e per l'estrazione di informazioni importanti	Sviluppato per i test sul web Utilizzato per l'automazione dei test delle applicazioni web Scraping di siti web ad alto contenuto di JavaScript Utilizzato per progetti complessi su piccola scala e di basso livello
Vantaggio	Un modo semplice per recuperare i dati da un URL Scraping di dati dal web Permette di leggere, scrivere, pubblicare, cancellare e aggiornare i dati per l'URL indicato. È molto facile gestire i cookie e le sessioni	Library portatile Funziona su Linux, Windows e Mac Una delle librerie di scraping più veloci Può estrarre siti web molto più velocemente di altri strumenti Consuma meno memoria e CPU Costruisce un'applicazione robusta e flessibile con diverse funzioni	È facile imparare a usarlo Il supporto della community è rapido e aiuta a risolvere i problemi.	Gestisce il sito web Core JavaScript-heavy Può gestire richieste AJAX e PJAX

Selezionatori	Nessuno	JCSS e XPath	CSS	CSS e Xpath
Documentazione	Dettagliata e semplice da capire	Dettagliata e semplice da capire	Dettagliata e semplice da capire	Dettagliata e molto complessa
Stelle di GitHub	46.8 k	42.7 k	–	22.7 k
Referenza	Chandra e Varanasi [27]	Kouzis-Loukas [28]	Richardson [29]	Sharma [30]

Tabella 2. Confronto tra quattro strumenti di text mining.

2.2.2 Limitazione dell'accesso

Quando un web crawler visita un sito web, potrebbe trovare restrizioni d'accesso in alcune pagine o nell'intero sito. Queste restrizioni sono implementate principalmente dai proprietari del sito per motivi di riservatezza, integrità e qualità dei dati, oltre che per questioni legali. Un crawler di solito esegue più richieste al secondo e scarica file di grandi dimensioni per ottenere i dati in breve tempo, il che può causare il blocco del server di un sito web. Per affrontare questo problema, sono disponibili numerosi metodi. Il tag Canonical, il robots.txt, l'x-robots-tag, il tag meta-robots e altri sono file usati dai proprietari dei siti web per dare delle istruzioni di scraping del sito senza che si creino problemi. Ad esempio, i file "robots.txt" sono spesso utilizzati dai siti web per comunicare le loro intenzioni di scraping e crawling. I file robots.txt consentono ai bot di scraping di effettuare il crawling di siti specifici, mentre i bot malevoli, d'altro canto, non sono interessati ai file robots.txt (che agiscono come un segnale di "non entrare"), come spiegato di seguito nella Figura 3.

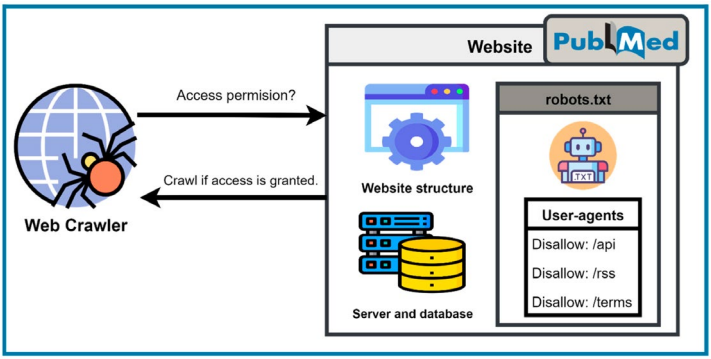


Figura 3. Schema delle restrizioni di accesso.

2.2.3 Raccolta di dati da diverse fonti

Il ritmo con cui vengono generati i dati medici aumenta di giorno in giorno nell'anno dell'esplosione massiccia delle informazioni e le informazioni globali vengono prodotte in grandi quantità in ogni campo, compreso quello sanitario [31,32]. I dati amministrativi, i dati biometrici, le registrazioni cliniche, le diagnosi, le radiografie, le cartelle cliniche elettroniche, i dati dei rapporti dei pazienti, i trattamenti, i risultati e altri tipi di dati sono tutti inclusi nei dati medici. Queste caratteristiche numerose e complesse rendono i dati difficili da trattare per ottenere risultati significativi e inediti. I centri sanitari e le istituzioni mediche di tutto il mondo hanno proposto una serie di sistemi informativi medici per gestire i dati in rapida crescita e fornire i migliori servizi e cure possibili ai pazienti [32]. Il modo più comune per raccogliere e archiviare i dati è il software di gestione, che può archiviare tutti i record elettronici e non elettronici. Sono disponibili diversi prodotti software, ad esempio eHospital Systems (adroitinfosystems.com/products/ehospital-systems, visitato l'11 aprile 2022) e DocPulse Clinic/Hospital Information Management System (docpulse.com, visitato l'11 aprile 2022).

La raccolta dei dati dalle fonti è il passo fondamentale per il text mining. Nella scienza medica vengono generati rapidamente

vari tipi di dati medici e tendenze, che possono essere distinti in cinque categorie:

- Software di gestione ospedaliera (dati dei pazienti/narrazioni cliniche).
- Studi clinici.
- Dati di ricerca in medicina.
- Piattaforme di pubblicazione per la medicina (PubMed, per esempio).
- Prodotti farmaceutici e dati normativi.

Le tabelle 3, 4 e 5 forniscono ulteriori dettagli sui diversi tipi di fonti di dati. I dati dei pazienti generati dalle sperimentazioni cliniche sono disponibili da diverse fonti, come mostrato nella Tabella 3. I ricercatori medici traggono vantaggio dai database ad open source perché dispongono di enormi volumi di dati di grande qualità, con un'ampia copertura e che consentono una strategia di studio economicamente vantaggiosa. Esistono numerosi set di dati e database disponibili pubblicamente relativi a vari settori della medicina, che presentano molti dettagli sulle cartelle cliniche (Tabella 4). Le informazioni testuali sono in rapida crescita ed è difficile raccogliere dati concisi in modo rapido e strutturato. La letteratura pubblicata è la fonte più ampia e primaria di informazioni testuali in ambito sanitario (Tabella 5).

Banche dati/registri	Numeri del trial	Fornito da	Posizione	Anno di fondazione	URL
ClinicalTrials.gov	405,612	Biblioteca nazionale di medicina degli Stati Uniti	Bethesda, MD, USA	1997	https://clinicaltrials.gov/ (consultato l'11 aprile 2022)
Registro centrale Cochrane degli studi controllati (CENTRAL)	1,854,672	un componente della Cochrane Library	Londra, Regno Unito	1996	https://www.cochranelibrary.com/central (consultato l'11 aprile 2022)

Piattaforma del registro internazionale degli studi clinici dell'OMS (ICTRP)	353,502	Organizzazione Mondiale della Sanità	Ginevra, Svizzera	–	https://trialsearch.who.int/ (consultato l'11 aprile 2022)
Database degli studi clinici dell'Unione Europea	60,321	Agenzia europea per i medicinali	Amsterdam, Paesi Bassi	2004	https://www.clinicaltrialsregister.eu/ctr-search/search (consultato l'11 aprile 2022)
CenterWatch	50,112	–	Boston, MA, USA	1994	http://www.centerwatch.com/clinical-trials/listings/ (consultato l'11 aprile 2022)
Registro tedesco degli studi clinici (Deutsches Register Klinischer Studien-DRKS)	>13,000	Istituto federale per i farmaci e i dispositivi medici	Colonia, Germania	–	https://www.bfarm.de/EN/BfArM/Tasks/German-Clinical-Trials-Register/_node.html (consultato l'11 aprile 2022)

Tabella 3. Banche dati e registri per gli studi clinici.

Banche dati	Numero di set di dati	Di proprietà di	Domini	Risorse disponibili	URL	Rif.
Centro di coordinamento delle informazioni sui campioni biologici e sui depositi di dati (BioLINCC)	262	Istituto Nazionale della Sanità, Calverton, MD, USA	Cardiovascolare, polmonare ed ematologico	Campioni e set di dati dello studio	https://biolincc.nhlbi.nih.gov/studies/ (consultato il 4 aprile 2022)	[33]
Sistema informativo per la ricerca biomedica traslazionale (BTRIS)	Cinque miliardi di righe di dati	Bethesda, MD, USA	Più soggetti	Set di dati dello studio	https://btris.nih.gov/ (consultato il 4 aprile 2022)	[34]
Richiesta di studio di dati clinici	3135	L'associazione dei finanziatori di studi clinici	Più soggetti	Set di dati dello studio	https://www.clinicalstudydatarequest.com/ (consultato il 4 aprile 2022)	[35]

Sorveglianza, epidemiologia e risultati finali (SEER)	–	Istituto Nazionale del Cancro, Bethesda, MD, USA	Cancro (tutti i tipi) - Stadio e dettagli istologici	Set di dati dello studio	https://seer.cancer.gov/ (consultato il 4 aprile 2022)	[36]
Informazioni mediche Mart per la terapia intensiva (MIMIC) MIMIC-III	53.423 pazienti	Laboratorio di fisiologia computazionale del MIT, Cambridge, MA, USA	Terapia intensiva	Dati del paziente (segni vitali, farmaci, misurazioni di laboratorio, osservazioni e note registrate dagli operatori sanitari, dati di sopravvivenza, durata della degenza, referti di diagnostica per immagini, codici diagnostici, codici di procedura e bilancio dei fluidi).	https://mimic.mit.edu/ (consultato il 4 aprile 2022)	[37,38]
MIMIC-CXR	65.379 pazienti (377.110 immagini di radiografie del torace)					[39]
Indagine nazionale sulla salute e la nutrizione (NHANES)	–	Centri per il controllo e la prevenzione delle malattie, Hyattsville, MD, USA	Valutazione della dieta e altri tipi di sorveglianza nutrizionale	dati sullo stato nutrizionale, sull'alimentazione, sulle misure antropometriche, sugli esami di laboratorio, sui campioni biologici e sui risultati clinici.	https://www.cdc.gov/nchs/nhanes/index.htm (consultato il 4 aprile 2022)	[40]
Onere globale delle malattie (GBDx)	–	Istituto per la metrica e la valutazione della salute, Seattle, WA, USA	Modelli epidemici e carico di malattia	Indagini, censimenti, parametri vitali e altri dati relativi alla salute	https://ghdx.healthdata.org/ (consultato il 4 aprile 2022)	[41]
Biobanca del Regno Unito (UKB)	0,5 milioni di euro	Stockport, Regno Unito	Informazioni genetiche e sanitarie approfondite	Dati genetici, biospecifici e sanitari	https://www.ukbiobank.ac.uk/ (consultato il 4 aprile 2022)	[42]

L'Atlante del genoma del cancro (TCGA)	caratterizzato molecolarmente oltre 20.000 campioni di cancro che coprono 33 tipi di tumore	Istituto Nazionale del Cancro, NIH, Bethesda, MD, USA	Genomica del cancro	oltre 2,5 petabyte di dati epigenomici, proteomici, trascrittomici e genomici	https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (consultato il 4 aprile 2022)	[43]
Omnibus dell'espressione genica (GEO)	4.981.280 campioni	Centro nazionale di bioinformatica (NCBI), NIH, Bethesda, MD, USA	Sequenziamento ed espressione genica	4348 set di dati disponibili	https://www.ncbi.nlm.nih.gov/geo/ (consultato il 4 aprile 2022)	[44]

Tabella 4. Dati della ricerca in Medicina.

Fonte	Articoli (milioni)	Lanciato da	Tipo di pubblicazione	Argomento	In linea	Collegamento
PubMed	33	Centro nazionale per le informazioni biotecnologiche (NCBI)	Estratti	Scienze biomediche e della vita	1996	https://www.ncbi.nlm.nih.gov/pubmed/ (consultato il 4 aprile 2022)
PubMed Central (PMC)	7.6	Centro nazionale per le informazioni biotecnologiche (NCBI)	Testo completo	Scienze biomediche e della vita	2000	https://www.ncbi.nlm.nih.gov/pmc/ (consultato il 4 aprile 2022)
Biblioteca Cochrane	–	Cochrane	Estratti e testo completo	Assistenza sanitaria	–	https://www.cochranelibrary.com/search (consultato il 4 aprile 2022)
bioRxiv	–	Laboratorio Cold Spring Harbor (CSHL)	Preparazioni inedite	Scienze biologiche	2013	https://www.biorxiv.org/ (consultato il 4 aprile 2022)
medRxiv	–	Laboratorio Cold Spring Harbor (CSHL)	Manoscritti inediti	Scienze della salute	2019	https://www.medrxiv.org/ (consultato il 4 aprile 2022)

arXiv	2.05	Cornell Tech	Non sottoposto a revisione inter pares	Multidisciplinare	1991	https://arxiv.org/ (consultato il 4 aprile 2022)
Google Scholar	100 (nel 2014)	Google	testo completo o metadati	Multidisciplinare	2004	https://scholar.google.com/ (consultato il 4 aprile 2022)
Studioso della Semantica	205.25	Istituto Allen per Intelligenza artificiale	Estratti e testo completo	Multidisciplinare	2015	https://www.semanticscholar.org/ (consultato il 4 aprile 2022)
Elsevier	17 (dal 2018)	Elsevier	Estratti e testo completo	Multidisciplinare	1880	https://www.elsevier.com/ (consultato il 4 aprile 2022)
Springer Nature	–	Gruppo Springer Nature	Estratti e testo completo	Multidisciplinare	2015	https://www.springernature.com/ (consultato il 4 aprile 2022)
Springer	–	Springer Nature	Estratti e testo completo	Multidisciplinare	1842	https://link.springer.com/ (consultato il 4 aprile 2022)

Tabella 5. Fonti di letteratura biomedica.

2.3 PREPARAZIONE DEI DATI

Nella terza fase (preparazione dei dati) di CRISP-DM, viene creato un set di dati finale dai dati grezzi, che verrà utilizzato nello strumento di modellazione. Questa fase rappresenta la parte principale (circa l'80%) di un progetto di text/data mining. I passaggi della fase di preparazione dei dati sono (1) la selezione dei dati (per scegliere il set di dati e i relativi attributi da utilizzare per l'analisi in base agli obiettivi del progetto, alla qualità, al tipo e al volume dei dati), (2) la pulizia dei dati (per stimare i dati mancanti e migliorare il set di dati correggendo, notificando o rimuovendo i valori errati), (3) costruzione dei dati (per creare attributi derivati o record completamente nuovi, nonché per trasformare i dati come necessario), (4) integrazione dei dati (per creare nuovi set di dati e aggregare nuovi valori combinando i dati da più fonti), (5) formazione dei dati (per rimuovere i carat-

teri inappropriati dai dati e modificare il formato o il design dei dati in modo che si adattino al modello) [14].

2.3.1 Pulizia dei dati/trasformazione dei dati

L'obiettivo principale della pulizia dei dati è individuare e rimuovere i dati duplicati e gli errori da un set di dati per creare un set di dati affidabile. La pulizia dei dati comporta l'identificazione e la rimozione di voci da un set di dati che sono corrotte, errate, duplicate, incomplete o formattate in modo improprio (vedi Figura 4). La pulizia dei dati è necessaria per analizzare le informazioni provenienti da più fonti [45,46,47].

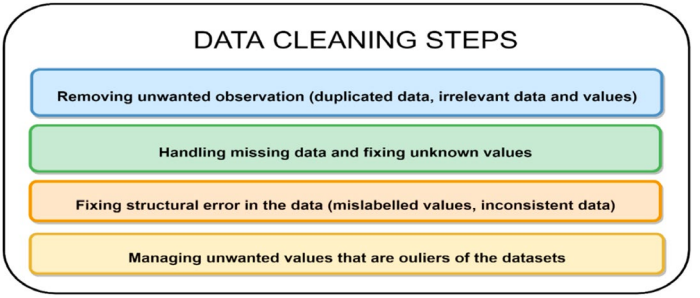


Figura 4. Fasi della pulizia dei dati.

Nelle sezioni seguenti vengono discussi vari strumenti e librerie python correlate.

Le librerie Python per la pulizia dei dati sono:

- NumPy: è una libreria Python open-source rapida e facile da usare per l'elaborazione dei dati. Poiché molte delle più note librerie Python, tra cui Pandas e Matplotlib, sono basate su NumPy, si tratta di una libreria fondamentale per l'ambiente della scienza dei dati. Lo scopo principale della libreria NumPy è la manipolazione diretta di grandi array multidimensionali, vettori e matrici. Per i calcoli numerici, NumPy offre anche funzioni efficacemente implementate [48].

- Le attività di elaborazione dei dati, come la pulizia, la manipolazione e l'analisi dei dati, vengono eseguite utilizzando la nota libreria Python Pandas. La libreria Python per l'analisi dei dati verrà qui chiamata "Pandas". Nella libreria sono disponibili diversi moduli per la lettura, l'elaborazione e la scrittura di file CSV, JSON ed Excel. Sebbene esistano molti strumenti di pulizia dei dati, la gestione e l'esplorazione dei dati con la libreria Pandas è incredibilmente rapida ed efficace [49].
- Una libreria Python open-source per automatizzare le procedure di pulizia dei dati si chiama DataCleaner. Le funzioni di preelaborazione dei dati di Pandas Dataframe e scikit-learn comprendono due moduli separati [50].

Dopo la pulizia, i dati vengono trasformati nel formato corretto (Excel, JSON o XML). La trasformazione dei dati semplifica la pre-elaborazione dei dati e/o del testo. A seconda delle modifiche da apportare, la trasformazione dei dati può essere semplice o complicata. Dopo la trasformazione, i dati sono più facili da usare sia per l'uomo che per il computer, perché sono più strutturati e organizzati. Inoltre, diventano più semplici da integrare in vari programmi e sistemi [46].

Nelle sezioni seguenti vengono discussi vari strumenti correlati.

1. Generation of Bibliographic Data è noto come GROBID. Si tratta di una libreria di apprendimento automatico che si è trasformata in una libreria open-source all'avanguardia per l'estrazione di metadati da documenti tecnici e scientifici in formato PDF. La libreria prevede di ricostruire la struttura logica del documento originale, oltre alla semplice estrazione bibliografica, al fine di supportare non solo dei processi da libreria digitale avanzata su larga scala, ma anche l'analisi del testo. Per questo motivo GROBID sviluppa soluzioni completamente automatizzate basate su modelli di apprendimento automatico. ResearchGate, Mendeley, CERN Inspire e HAL, l'archivio nazionale francese delle

pubblicazioni, sono solo alcuni dei servizi scientifici commerciali e open access a cui la biblioteca è collegata. Il risultato è quello di estrarre e trasformare i documenti PDF in formato XML TEI, integrare le informazioni estratte con altri servizi online e riunire i risultati nei documenti PDF di articoli scientifici [51,52].

2. BioC è un formato semplice e diretto per lo scambio di dati e annotazioni testuali e per la semplice elaborazione del testo. Il suo obiettivo principale è quello di fornire una grande quantità di dati e articoli di ricerca per il text mining e il recupero di informazioni. Sono disponibili in diversi formati di file, tra cui BioC XML, BioC JSON, Unicode e ASCII. Questi formati sono disponibili tramite API Web o FTP [53].

In sintesi, la pulizia dei dati migliora la consistenza di un set di dati, mentre la trasformazione semplifica l'elaborazione dei dati. Entrambi i processi migliorano la qualità del set di dati di addestramento per la costruzione del modello.

2.3.2 Ingegneria delle caratteristiche

La scelta, la modifica e la conversione dei dati grezzi in caratteristiche che possono essere utilizzate nell'apprendimento supervisionato è un processo di ingegneria delle caratteristiche, spesso definito estrazione di caratteristiche. Questa tecnica di apprendimento automatico, l'ingegneria delle caratteristiche, utilizza i dati per generare nuove variabili non presenti nel set di addestramento. Per semplificare e accelerare le trasformazioni dei dati, migliorando al contempo l'accuratezza del modello, può generare nuove caratteristiche per l'apprendimento supervisionato e non supervisionato. Con i modelli di apprendimento automatico, l'ingegneria delle caratteristiche è necessaria. Indipendentemente dall'architettura o dai dati, una caratteristica errata influisce direttamente sul modello. Sono disponibili numerosi strumenti per automatizzare l'intero processo di ingegnerizza-

zione delle caratteristiche e per generare molte caratteristiche in breve tempo per compiti di classificazione e regressione. Alcuni strumenti di feature engineering sono FeatureTools, AutoFeat, TsFresh, Turi, Azure Machine Learning Studio, ZOMBIE, FeatureFu e OneBM [54,55].

Vijithananda et al. [56] hanno estratto caratteristiche da immagini di risonanza magnetica ADC di un tumore cerebrale. Le seguenti caratteristiche sono state estratte da fette di immagini etichettate MRI ADC del cervello di 195 pazienti: obliquità, tonalità dei cluster, valori dei pixel (dati demografici), prominenza, caratteristiche della matrice di co-occorrenza a livello grigio (GLCM [*Grey Level Co-occurrence Matrix*]), energia, contrasto, entropia, varianza, media, correlazione, omogeneità e curtosi. Sia l'omogeneità che l'obliquità di GLCM sono state escluse perché hanno ottenuto i punteggi più bassi nel processo di selezione delle caratteristiche dell'ANOVA f-test. Il classificatore Random Forest ha superato le prestazioni di alberi decisionali, Nave Bayes, analisi discriminante lineare, K-Nearest Neighbors (KNN) e regressione logistica, ed è stato scelto per l'ulteriore sviluppo del modello. Il modello finale ha ottenuto un'accuratezza del 90,41% nella previsione di neoplasie maligne e benigne.

2.3.3 Ricerca di parole chiave

L'estrazione di parole o frasi chiave da documenti di testo è nota come estrazione di parole chiave. Esse vengono scelte tra le frasi presenti nel documento di testo e descrivono l'argomento del documento. Ci sono molti metodi comunemente usati per l'estrazione automatica delle parole chiave. Vengono utilizzati nei processi di estrazione automatica delle parole chiave dai documenti per selezionare le parole o le frasi più frequenti e significative dal documento di testo. Ciò classifica i metodi di estrazione delle parole chiave come parte del campo dell'elaborazione del linguaggio naturale, importante per l'apprendimento automatico e l'intelligenza artificiale. [57]. Gli estrattori di paro-

le chiave vengono utilizzati per estrarre parole (parole chiave) o gruppi di due o più parole che formano una frase (frasi chiave).

FlashText, ad esempio, è un pacchetto Python gratuito e open-source che consente la ricerca e la sostituzione di parole chiave ed è uno degli strumenti di estrazione di parole chiave di cui abbiamo appena parlato [58]. FlashText esegue un'analisi completa utilizzando un algoritmo di Aho-Corasick e un dizionario Trie. In generale, la corrispondenza delle parole chiave prevede la scansione del corpus (i documenti creati dall'uomo sono un ampio insieme strutturato di testi) per ogni termine. Si consideri il seguente scenario: una persona ha 100 parole chiave e deve cercare in 2000 documenti; si seleziona un singolo termine alla volta e si esegue una ricerca nel corpus di 2k; la ricerca continua per 100 × 2000, cioè 200.000 iterazioni. Oltre a questo strumento per la ricerca di parole chiave, sono stati selezionati quattro strumenti basati su Python tra i vari strumenti per l'estrazione di parole e frasi disponibili; le loro caratteristiche, i vantaggi e i compiti NLP sono riportati nella Tabella 6.

	Toolkit per il linguaggio naturale	**SpaCy**	**Kit di strumenti NLP Scikit-Learn**	**Gensim**
Che cos'è?	piattaforma Python open-source per la gestione dei dati sul linguaggio umano	libreria Python open-source per l'elaborazione avanzata del linguaggio naturale	libreria software di apprendimento automatico per il linguaggio di programmazione Python	la libreria Python più veloce per addestrare l'incorporazione dei vettori
Caratteristiche			Basato su NumPy, SciPy e Matplotlib Un modo semplice ed efficiente per analizzare i dati predittivi Facilmente accessibile e riutilizzabile in diversi contesti	

Vantaggio	Le più note e complete librerie NLP con numerose estensioni offre supporto in molte più lingue rispetto agli altri	facile da usare completamente integrato con Python compatibile con altri framework di deep learning sono disponibili molti modelli statistici già addestrati applicabile a molte lingue diverse velocità e prestazioni elevate disponibile gratuitamente in grado di elaborare testi lunghi utilizzabile indipendentemente dalla piattaforma	strumenti semplici ed efficienti per l'apprendimento automatico, il data mining e l'analisi dei dati gratis per tutti applicabile a diverse aree applicative, come l'elaborazione del linguaggio naturale	Fornisce modelli e corpora pronti all'uso. Modelli preaddestrati per aree specifiche come l'assistenza sanitaria Elabora grandi quantità di dati utilizzando lo streaming dei dati
Compiti di PNL	Classificazione Tokenizzazione Stemming Etichettatura Parsing	Classificazione Tokenizzazione Stemming Etichettatura Parsing Riconoscimento di entità denominate Sentiment analysis	Classificazione Modellazione degli argomenti Sentiment analysis	Somiglianza del testo Riassunto del testo Modellazione degli argomenti
Stelle di GitHub	10.4 k	22.4 k	49 k	12.9 k

Sito web	nltk.org (visitato il 16 marzo 2022)	spacy.io (visitato il 16 marzo 2022)	scikit-learn.org (visitato il 16 marzo 2022)	radimrehurek.com/gensim/ (consultato il 16 marzo 2022)
Referenza	Bird et al. [59]	Honnibal [60]	Pedregosa et al. [61], Pinto et al. [62]	Rehurek e Sojka [63]

Tabella 6. Ricerca di contenuti rilevanti.

2.4 MODELLAZIONE

Nella quarta fase (nota come modellazione) del CRISP-DM, vengono testate e calibrate varie tecniche di modellazione regolando i parametri del modello per ottenere i risultati migliori [14]. Le fasi del processo di modellazione sono: (1) la scelta di una tecnica di modellazione per selezionare uno o più modelli/algoritmi/ipotesi specifici per l'attività, (2) la creazione di test per determinare la forza del modello quanto a qualità e validità, (3) la costruzione di modelli (utilizzare lo strumento di modellazione per costruire modelli dal set di dati preparato), (4) la valutazione dei modelli per spiegare il risultato del modello in base alla conoscenza del soggetto, alle norme di successo predeterminate e al progetto di test, classificare i modelli multipli generati e regolare nuovamente le impostazioni dei parametri, se necessario.

Tra i vari modelli disponibili per l'organizzazione e l'analisi dei dati, la scelta di un modello dipende dallo scopo (ad esempio, la previsione) e dal tipo di dati utilizzati (non strutturati o strutturati). Un modello è un insieme di dati, modelli e statistiche. I modelli di data-mining disponibili si dividono in due categorie: predittivi e descrittivi. I modelli descrittivi sono spesso utilizzati per determinare modelli nei dati che possono essere spiegati dall'uomo. I modelli predittivi utilizzano i risultati noti di vari set di dati per prevedere valori futuri o non identificati di altre variabili di interesse. I modelli predittivi si basano solitamente sui dati forniti in precedenza e sui loro risultati. Classificazione, previsione, regressione e analisi delle serie temporali sono compiti dei modelli predittivi. I compiti di data mining dei modelli descrittivi comprendono il raggruppamento, l'asso-

ciazione di regole, la scoperta di sequenze e la sintesi (Figura 5). Per la previsione e l'analisi dei modelli nei dati sono disponibili diversi algoritmi/metodi. Tuttavia, la scelta dell'algoritmo dipende principalmente dalle variabili dipendenti, etichettate o non etichettate. Se le variabili dipendenti nel set di dati sono etichettate, si utilizza un algoritmo di apprendimento supervisionato. Gli algoritmi comunemente usati sono quelli degli alberi decisionali, i Random Forest (RF), le macchine a vettori di supporto (SVM) e il modello di rischio competitivo. Al contrario, se le variabili dipendenti dei dati non sono etichettate, si utilizza un metodo di apprendimento non supervisionato. L'analisi di clustering, il clustering di partizione, il clustering gerarchico, l'analisi delle componenti principali (PCA) e l'analisi di associazione sono alcuni degli algoritmi di apprendimento non supervisionato [64,65].

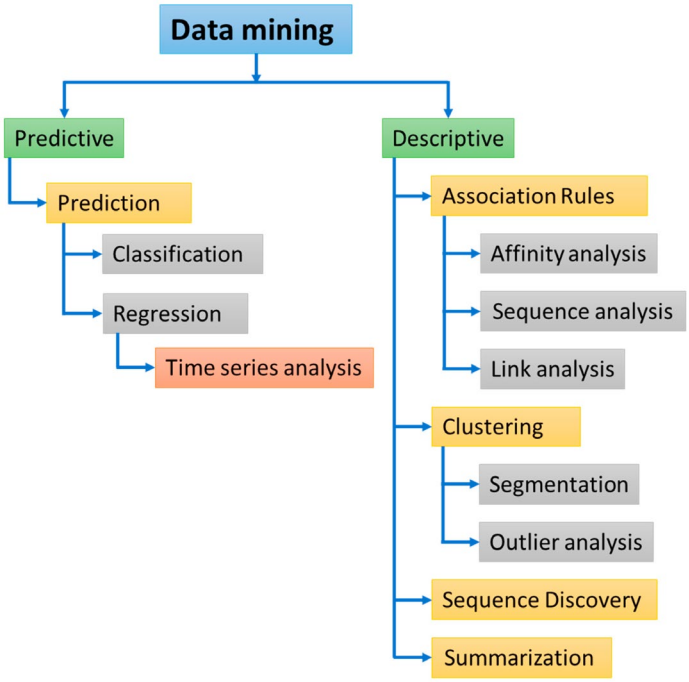


Figura 5. Attività di data mining predittivo e descrittivo.

Il set di dati è la principale differenza tra apprendimento automatico supervisionato e non supervisionato. Si parla di apprendimento supervisionato se si utilizza un set di dati etichettati per l'input e l'output, mentre le tecniche di apprendimento non supervisionato utilizzano dati non etichettati. Come suggerisce il nome, l'apprendimento supervisionato prevede che l'addestramento di un modello avvenga sotto una supervisione esterna. L'apprendimento non supervisionato, invece, non prevede alcuna supervisione. Inoltre, nel caso dell'apprendimento supervisionato, l'obiettivo è prevedere il risultato che daranno dei nuovi dati. Nel caso dell'apprendimento non supervisionato, l'obiettivo è trovare schemi nascosti e ricavare informazioni da enormi quantità di nuovi dati. A differenza dei modelli di apprendimento supervisionato, che sono semplici, i modelli di apprendimento non supervisionato richiedono un ampio set di addestramento per produrre i risultati desiderati, il che li rende complessi a livello computazionale. Alcune delle applicazioni dei modelli di apprendimento supervisionato comprendono la diagnosi, il rilevamento delle frodi di identità, la classificazione delle immagini, la previsione dei prezzi, la sentiment analysis, il rilevamento dello spam, le previsioni di mercato e le previsioni meteorologiche. I modelli di apprendimento non supervisionato sono utilizzati nelle pipeline per il rilevamento di anomalie, la visualizzazione di big data, le buyer personas, lo stimolo di alcune caratteristiche, i sistemi raccomandati, la scoperta di strutture e il marketing mirato [64,66].

Come esempio di modellazione, Zowalla et al. hanno studiato l'idoneità di un WebCrawler (StormCrawler) per l'acquisizione di tutti i contenuti web relativi alla salute sul Web, in tedesco (Germania, Austria e Svizzera) [67]. A tal fine, è stato addestrato un modello di classificatore a macchina vettoriale di supporto per distinguere tra pagine web relative alla salute e non relative alla salute, utilizzando il set di dati creato dal web tedesco. Il modello è stato testato in termini di accuratezza e precisione su una divisione 80/20 tra training e test e rispetto a un set di dati convalidato in massa. Per la previsione delle malattie cardiova-

scolari, la tecnica più adatta è stata l'"albero decisionale" rispetto ad altre otto tecniche, ossia Deep Learning, Nearest Neighbor (k-NN), Gradient Boosted Tree, modello lineare generalizzato, regressione logistica, Naïve Bayes, Random Forest e induzione alla regola [18]. Inoltre, alcuni parametri sono stati ottimizzati utilizzando l'operatore apposito, per ottenere risultati migliori quando si utilizzava l'albero decisionale.

2.5 CONVALIDA E TEST DEL MODELLO DI DATI

L'obiettivo principale di questa fase è quello di convalidare e testare il modello selezionato per i dati dello sviluppo del modello. La procedura di validazione serve a garantire che il modello sviluppato sia sufficientemente accurato per l'uso previsto [68]. La prima parte di questa fase, la convalida del modello, è importante perché non si può fare affidamento sul modello usato o appena sviluppato solo perché è stato progettato per adattarsi ai dati di addestramento e dimostrare che i dati di addestramento si adattano bene al modello. Per convalidare un modello, le previsioni di output vengono effettuate in scenari non correlati al set di addestramento e vengono calcolate le stesse misure statistiche di adattamento. La seconda metà di questa fase consiste nel testare il modello con i dati di prova e confrontare la sua accuratezza con i risultati della fase di validazione. Solo quando un modello viene confrontato con i dati di prova e i calcoli statistici mostrano una corrispondenza soddisfacente, viene considerato "pronto". Per la classificazione di campioni tumorali e non tumorali, Dong et al. [69] hanno utilizzato un set di dati di addestramento (che consiste in dati grezzi di spettrometria di massa (MS) ottenuti da 194 campioni tumorali e non tumorali accoppiati) per addestrare diversi modelli e hanno utilizzato un tipo simile di set di dati (che consiste in dati grezzi MS ottenuti da 58 campioni tumorali e non tumorali accoppiati) come set di dati di prova. Sono stati confrontati la rete neurale convoluzionale (CNN), l'albero decisionale gradient boosting (GBDT [*Gradient Boosting Decision Tree*]), la support-vector machine (SVM), l'ana-

lisi delle componenti principali (PCA) più SVM, la regressione logistica (LR) e il random forest (RF) e il modello CNN è quello che ha mostrato la maggiore accuratezza. Alcuni degli strumenti di test per la validazione dei modelli ML includono Apache Spark, Excel, Hadoop, KNIME, Python, R, RapidMiner, SAS, SQL e Tableau.

2.6 VALUTAZIONE

Nella quinta fase (nota come valutazione) del CRISP-DM, viene condotta una valutazione più approfondita e una revisione del modello per garantire che raggiunga gli obiettivi aziendali. I passaggi della fase di valutazione sono: (1) la valutazione dei risultati per osservare il grado di raggiungimento degli obiettivi del progetto da parte del modello, scoprire ulteriori vincoli, informazioni o indizi sui percorsi futuri e presentare la dichiarazione finale del progetto; (2) il processo di revisione per condurre un esame più approfondito del progetto e affrontare i problemi di garanzia della qualità; (3) la decisione di ulteriori passi per determinare se procedere o meno con l'implementazione o apportare modifiche per il miglioramento [14].

Dopo l'analisi dei dati testuali, il passo successivo è quello di visualizzarli in modo significativo ai fini dell'interpretazione e della comunicazione. La visualizzazione del testo avviene principalmente attraverso l'uso di grafici, diagrammi, mappe, linee del tempo, reti, cloud di parole e così via. I risultati così visualizzati permettono all'uomo di leggere gli aspetti più importanti di una grande quantità di informazioni. Sono disponibili diversi strumenti per visualizzare i dati analizzati. Questi strumenti facilitano l'identificazione e la scoperta di modelli, anomalie, tendenze e intuizioni nei dati in modo diretto e comprensibile. Una visualizzazione efficace dei dati presenta vantaggi e benefici quali la facile comprensione dei risultati, un processo decisionale rapido e senza sforzo e un maggior grado di coinvolgimento di un pubblico eterogeneo rispetto ad altri metodi di comunicazione (ad esempio, la comunicazione verbale). Per una visualizzazione

dei dati di successo, esistono tre regole principali: (1) bisogna selezionare lo stile di visualizzazione appropriato allo scopo, (2) lo stile di visualizzazione selezionato deve essere appropriato per il pubblico a cui è rivolto e (3) lo stile di visualizzazione scelto deve essere accompagnato da un design grafico efficace [70]. Gli aspetti più importanti della selezione dello stile di visualizzazione appropriato sono la considerazione dei dati selezionati e l'obiettivo della visualizzazione. Ad esempio, i grafici a linee e a barre sono adatti per confrontare i punti di dati in un set di dati. Per creare informazioni visive coinvolgenti ed efficaci sono disponibili diversi stili di visualizzazione: visualizzazione tipografica (ad esempio, cloud di parole), visualizzazione di diagrammi (ad esempio, ad albero), visualizzazione di grafici (ad esempio, grafico a barre/linee), visualizzazione 3D, ecc. Di seguito, nella Tabella 7, viene fornito un elenco dei vari stili di visualizzazione e alcuni degli strumenti disponibili per ciascuna categoria.

Stile di visualizzazione	Strumento [Riferimento]
Marcatura/illuminazione del testo	cite2vec [71], TopicLens [72], SurVis [73], Poemage [74], Overview [75]
Tag o cloud di parole	SentenTree [76], InfoVis [77], VisOHC [78], IncreSTS [79], Word storms [80]
Grafici a barre	TextTile [81], SentiCompass [82], NewsViews [83], WeiboEvents [84], CatStream [85]
Grafico di dispersione	PhenoLines [86], SocialBrands [87], TopicPanorama [88], #FluxFlow [89], PEARL [90]
Grafico a linee	Vispubdata.org [91], GameFlow [92], MultiConVis [93], Contextifier [94], Google+Ripples [95]
Collegamento al nodo	NEREx [96], iForum [97], NameClarifier [98], DIA2 [99], Information Cartography [100]
Albero	OpinionFlow [101], Rule-based Visual Mappings [102], HierarchicalTopics [103], Whisper [104], The World's Languages Explorer [105]
Matrice	Risoluzione interattiva dell'ambiguità [106], Matrici di impronte digitali [107], Pattern di ricorrenza concettuale [108], The Deshredder [109], Termite [110]

Linea temporale del grafico del flusso	VAiRoma [111], CiteRivers [112], ThemeDelta [113], EvoRiver [114], LeadLine [115]
Cronologia del flusso	TimeLineCurator [116], Profilazione visuale interattiva [117]
Visualizzazione radiale	ConToVi [118], ConVis [119]
Visualizzazione 3D	Struttura a due fasi [120]
Mappe/Geo chart	Twitter può salvare vite umane? [121], Visualizzazione di dati dinamici con le mappe [122], Rilevamento di anomalie spazio-temporali [123]

Tabella 7. Stile di visualizzazione dei dati con strumenti correlati.

Oltre a questi strumenti, sono disponibili software con enormi capacità di visualizzazione dei dati, come le tabelle pivot di Microsoft Excel, R, Tableau, Power-BI, datawrapper e Google Charts. Questi strumenti sono facili da usare e molto utili per creare una visualizzazione chiara e dinamica dei dati grazie alla loro interfaccia grafica interattiva. Inoltre, per la visualizzazione dei dati sono disponibili diverse librerie scritte in diversi linguaggi di programmazione, facili da usare per i programmatori, come le librerie JavaScript (ad esempio, D3.js, Chart.js e Highcharts), le librerie Python (ad esempio, Matplotlib, Seaborn e Plotly) e le librerie R (ad esempio, ggplot2, Leaflet ed Esquisse). Le principali sfide della visualizzazione dei dati sono l'enorme quantità di dati, la complessità dei dati e le voci mancanti/duplicate [124].

2.7 DISTRIBUZIONE

Nella fase di implementazione (sesta e ultima fase del CRISP-DM, Shearer [14]), le conoscenze acquisite dal progetto vengono organizzate e presentate (ad esempio, dimostrazioni dal vivo) in modo utile per il progetto, l'azienda e il cliente. La complessità di questa fase varia notevolmente. Le fasi della fase di implementazione sono le seguenti: (1) creare un piano di implementazione per formulare e annotare una strategia di implementazione del modello, (2) pianificare il monitoraggio e la manutenzione per

creare una pianificazione ben ponderata della manutenzione e del monitoraggio per evitare problemi durante la fase operativa di un modello, (3) produrre una relazione finale per preparare e presentare una relazione finale del progetto sotto forma di documento scritto e riunione verbale, e (4) rivedere il progetto per valutare i successi e gli insuccessi, nonché le potenziali aree di miglioramento per i progetti futuri.

3. CONCLUSIONI E PROSPETTIVE FUTURE

La quantità di dati di testi medici è in rapido aumento. Dai dati di testo medici, il data mining può essere utilizzato per estrarre informazioni o conoscenze nuove e utili. Il sistema CRISP-DM presentato in questo studio si concentra su ogni fase del data mining, utilizzando esempi medici per spiegare ogni fase. Gli autori hanno in programma di sviluppare un sistema di web crawling basato sull'intelligenza artificiale con visualizzazione 4D dei dati in modo sintetico e di facile comprensione e di utilizzare questi dati come fonte di informazioni per i ricercatori, oltre che per l'educazione dei pazienti e del personale medico.

3 - INTELLIGENZA ARTIFICIALE NELLE SCIENZE BIOLOGICHE

Tratto e tradotto da
Abhaya Bhardwaj, Shristi Kishore e Dhananjay K. Pandey,
Artificial Intelligence in Biological Sciences, Vita 2022, 12(9), 1430.

https://doi.org/10.3390/life12091430

Le parti omesse dal curatore rispetto all'originale sono indicate dal segno **[...]**

ABSTRACT

L'intelligenza artificiale (AI), attualmente un concetto all'avanguardia, può migliorare la qualità della vita degli esseri umani. I campi dell'IA e della ricerca biologica si stanno intrecciando sempre di più e i metodi per trovare e applicare le informazioni immagazzinate negli organismi viventi vengono costantemente perfezionati. Man mano che il campo dell'IA progredisce con algoritmi più addestrati, si amplia il potenziale della sua applicazione in epidemiologia, nello studio delle interazioni ospite-patogeno e nella progettazione di farmaci. L'IA viene ora applicata in diversi campi della scoperta di farmaci, della medicina personalizzata, dell'editing genico, della radiografia, dell'elaborazione delle immagini e della gestione dei farmaci. Nel prossimo futuro, grazie all'applicazione di tecnologie basate sull'IA, saranno possibili diagnosi più precise e trattamenti economicamente vantaggiosi. Nel campo dell'agricoltura, grazie all'applicazione di approcci avanzati basati sull'IA, gli agricoltori hanno ridotto gli sprechi, aumentato la produzione e diminuito il tempo necessario per portare i loro prodotti sul mercato. Inoltre, con

l'uso dell'IA attraverso programmi intelligenti basati sull'apprendimento automatico (ML) e sull'apprendimento profondo, è possibile modificare i percorsi metabolici dei sistemi viventi per ottenere i migliori risultati possibili con input ridotti. Tali sforzi possono migliorare i ceppi di specie microbiche per massimizzare la resa nel settore bio-industriale. Questo articolo riassume le potenzialità dell'IA e la sua applicazione a diversi campi della biologia, come la medicina, l'agricoltura e la bio-industria.

1. INTRODUZIONE

Finora non esiste una definizione precisa di intelligenza artificiale (IA), ma in generale ci si riferisce alla capacità di qualsiasi macchina di simulare l'intelligenza di organismi superiori. L'IA si radica in profondità in quasi tutte le branche della ricerca, tra cui filosofia, matematica, informatica, psicologia e biologia [1]. Un sistema di IA ideale sarebbe autoconsapevole, logico e in grado di imparare dall'esperienza. Sarebbe anche in grado di percepire l'ambiente esterno e di reagirvi. Con l'aiuto di algoritmi basati su approcci di machine learning (ML) e deep learning (DL), un tale sistema intelligente potrebbe essere sviluppato per svolgere attività che oggi richiedono l'intelletto umano [2]. Nel 1956 John McCarthy coniò per la prima volta il termine "intelligenza artificiale (AI)" per indicare un sistema di macchine intelligenti alla conferenza di Dartmouth [2]. Il primo lavoro significativo nell'ambito dell'IA è il contributo del matematico Alan Mathison Turing. Egli propose le sue idee in una conferenza pubblica a Londra sul concetto di autoapprendimento e di macchine autoistruite che imparano dalle proprie esperienze come fa un essere umano [3,4]. Grazie alla sua osservazione iniziale e alla concettualizzazione delle macchine intelligenti, Alan Turing è considerato dalla maggior parte delle persone il padre dell'IA e dell'informatica moderna. È stato uno dei primi sostenitori della teoria secondo cui il cervello umano funziona essenzialmente come un computer digitale [5]. Fu il pioniere dell'esperimento noto come

"Test di Turing", un momento cruciale nello sviluppo dell'IA (Figura 1). Il suo articolo, intitolato "*Computing Machinery and Intelligence*", esaminava la possibilità che un computer non vivente potesse pensare come un essere umano e rappresentava una pietra miliare in questo campo [3]. Altri eventi significativi hanno spianato la strada allo sviluppo dell'IA che vediamo oggi (Figura 2). Un programma di IA scritto da Arthur Samuel nel 1952 per il prototipo IBM 701 e un "topo virtuale" addestrato a muoversi attraverso un percorso predefinito basato su una rete neurale da John Holland sono stati lavori preliminari rivoluzionari [6,7]. Nel 1973, un gruppo di ingegneri giapponesi creò il primo robot umanoide, che all'epoca aveva diverse capacità notevoli per una macchina, tra cui la capacità di camminare in posizione eretta, di afferrare oggetti e di conversare in giapponese.

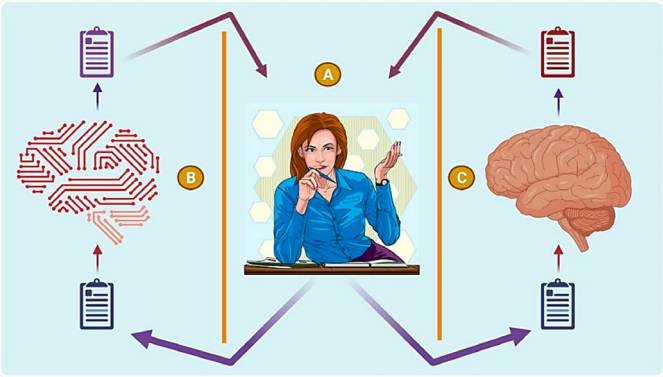


Figura 1. Alan Turing ideò il Test di Turing nel 1950. Il test prevede tre partecipanti, un interrogatore umano, una macchina intelligente e un altro umano, che possiamo chiamare A, B e C, rispettivamente. A non è a conoscenza dell'identità di B e C e può inviare e ricevere risposte solo sotto forma di messaggi di testo da B e C. A può porre a B e C una serie di domande e, in base alle loro risposte, se A non è in grado di distinguere quale tra B e C è un computer, allora il computer B può essere considerato intelligente e con capacità di pensiero. Se un interrogatore umano A non è in grado di trovare la differenza tra un altro umano e un computer, allora il computer è abbastanza intelligente da essere considerato umano. Questo test serve semplicemente a capire se una macchina ha o meno la capacità di pensare.

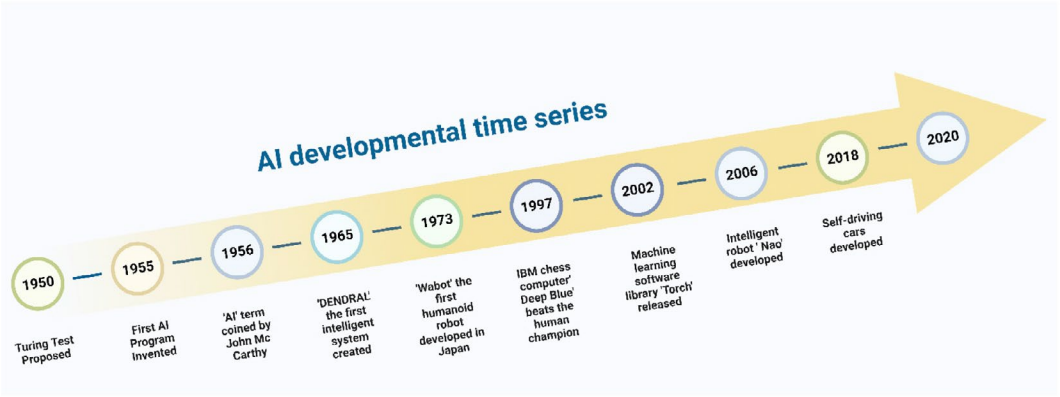


Figura 2. Linee temporali che evidenziano le importanti scoperte nel percorso evolutivo dell'intelligenza artificiale e la sua applicazione in vari campi.

Un altro evento significativo nella linea temporale dell'IA è stata la costruzione del supercomputer IBM, Deep Blue, che era in grado di giocare a scacchi in modo completamente indistinguibile dagli esseri umani. È stata la prima intelligenza artificiale a sconfiggere il Gran Maestro Garry Kasparov in una partita a tempo [8] (Figura 2). L'uso riuscito di approcci di pianificazione e percezione dell'IA può essere visto nei veicoli autonomi spaziali della NASA, che utilizzano la tecnologia per sterzare e muoversi da soli senza l'intervento umano [4]. L'apprendimento profondo e l'apprendimento automatico sono elementi cruciali dell'IA che si addestrano raccogliendo conoscenze da dati di varia origine, generati direttamente o indirettamente dal sistema di intelligenza naturale. Più questi algoritmi di deep learning e machine learning vengono addestrati utilizzando dati provenienti da varie fonti, più si potranno sviluppare sistemi artificiali avanzati, intelligenti e autoconsapevoli (Figura 3) [9].

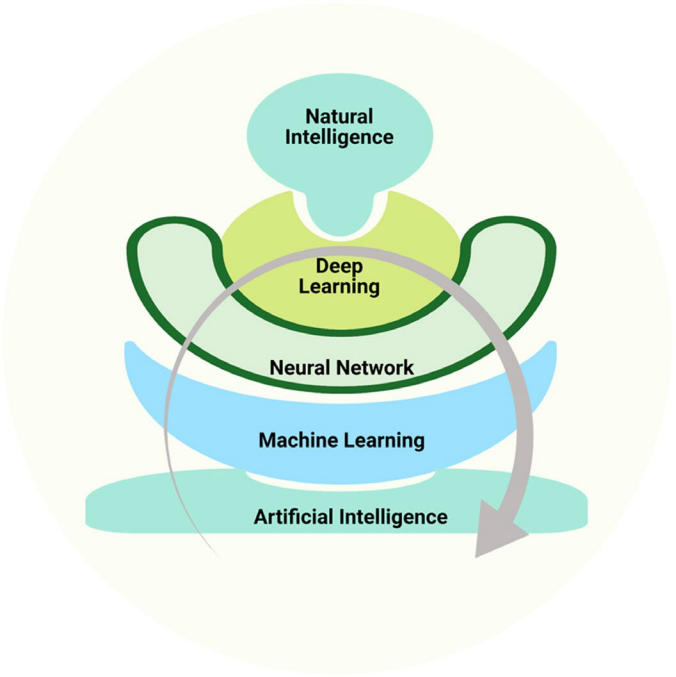


Figura 3. Rappresentazione schematica dei principali componenti dell'intelligenza artificiale e del processo di apprendimento continuo con l'aiuto dell'intelligenza naturale per creare macchine più intelligenti.

L'IA può essere classificata in due grandi categorie: l'IA ristretta o debole e l'Intelligenza Artificiale Generale o forte. L'IA debole tenta di copiare o imitare il pensiero cognitivo umano; consente di automatizzare la maggior parte dei compiti in modi che l'uomo non è in grado di fare [10]. Gli esempi più visibili di IA debole sono la funzione di pilota automatico di Tesla, il riconoscimento facciale dei nostri smartphone, il motore di ricerca di Google, l'IA di Instagram per la comprensione degli interessi degli utenti, Siri di Apple e Alexa di Amazon. L'IA forte è molto più avanzata e complessa dell'IA debole. L'IA forte non è limitata da leggi create dall'uomo e pensa e controlla il sistema in modo completamente autonomo. In parole povere, si tratta

di un programma o di una macchina che simula precise qualità cognitive o intellettuali umane, come le emozioni o delle grandi capacità di risoluzione dei problemi [11,12]. Un programma di IA debole è progettato per svolgere un solo compito alla volta, mentre un'IA forte è in grado di svolgere in modo efficiente numerosi compiti contemporaneamente [13]. Sebbene l'autoconsapevolezza sia la qualità più essenziale e unica che distingue l'IA forte dall'IA debole, essa è ancora nelle prime fasi di sviluppo e non ci sono applicazioni reali che possiamo osservare [13].

Le tecniche utilizzate nell'IA comprendono numerose varianti, ad esempio i sistemi basati su regole che si basano su rappresentazioni simboliche e lavorano su inferenze. I sistemi di IA sono basati sull'RNA che è progettato per lavorare sull'interfaccia con altri neuroni e sul peso delle connessioni [14]. Nonostante ciò, tutti questi sistemi condividono quattro caratteristiche. In primo luogo, rappresentano la conoscenza. I sistemi basati su regole, i sistemi basati su frame e le reti semantiche utilizzano una serie di regole se-allora, mentre le reti neurali artificiali utilizzano connessioni e pesi di connessione [15]. In secondo luogo, i sistemi ingegnerizzati di intelligenza artificiale sono in grado di apprendere. Come entità di autoapprendimento [16], raccolgono dati, ad esempio scegliendo i pesi di connessione appropriati per una rete neurale artificiale o definendo le regole per un sistema esperto basato su regole. In terzo luogo, hanno le regole, che possono essere implicite o esplicite in un sistema di intelligenza artificiale. La quarta è la ricerca, che può essere incorporata nel sistema in diversi modi. Ad esempio, può essere utilizzata per trovare gli stati che portano più rapidamente a una soluzione o per trovare il miglior insieme di pesi di connessione per una RNA minimizzando la funzione di fitness [1]. A seconda dell'algoritmo impiegato, l'IA può essere suddivisa in "basata su regole", nota anche come IA in termini generali, e "non basata su regole", nota anche come ML [*Machine Learning*]. Negli algoritmi basati su regole, vengono fornite istruzioni e ramificazioni condizionate per ottenere la soluzione migliore. Per esempio, l'algoritmo si atterrà completamente alle istruzioni e unirà i numeri quando il caso è definito

come "quando i numeri dei soggetti di due set di dati diversi sono uguali, devono essere trattati come duplicati e devono essere uniti". Un algoritmo basato su regole funziona bene quando le opzioni disponibili sono poche. Tuttavia, lo sviluppo di un algoritmo basato su regole è piuttosto impegnativo in scenari complessi. Il metodo ML, invece, sviluppa le regole direttamente dagli input di addestramento stabiliti e le implementa negli algoritmi ML attraverso metodi statistici. In questo modo, il ML si concentra sul riconoscimento rapido di modelli da un enorme volume di informazioni per fornire risultati più affidabili rispetto all'analisi e alle previsioni manuali [17].

L'IA ha fatto il suo ingresso nel campo biologico, dimostrando il suo valore attraverso procedure innovative e all'avanguardia [18]. Inoltre, il mondo ha assistito a una vera e propria rivoluzione nel campo delle tecnologie dell'informazione (IT), che ha portato alla produzione e all'archiviazione di un'enorme quantità di dati, non solo nel campo della tecnologia ma anche in altri settori negli ultimi anni. Sia l'informatica che la biologia sono fiorite nell'ultimo mezzo secolo. Secondo la legge di Moore, il numero di transistor su un chip raddoppia ogni due anni circa. È una conseguenza e un motore della rapida crescita della tecnologia dell'informazione [19]. Le risorse computazionali sono inestricabilmente legate ai big data, che comprendono informazioni annotate e grezze a causa del volume e della complessità sempre crescenti dei dati provenienti da più fonti [20,21]. Grazie agli sviluppi del sequenziamento e di altre tecniche high-throughput, negli ultimi anni le bioscienze e le industrie biotecnologiche hanno fatto passi da gigante [22]. Gli algoritmi basati sull'intelligenza artificiale hanno la capacità di memorizzare ed elaborare efficacemente grandi quantità di dati grezzi e non strutturati e di renderli disponibili per una rapida estrazione, necessaria per costruire un sistema informatico intelligente con capacità decisionali complesse [23,24]. Questi progressi nella generazione, archiviazione e analisi dei dati consentono lo sviluppo di un'ampia gamma di prodotti e servizi in diversi settori, tra cui le bioscienze [19]. Mentre i progressi nel campo dell'informatica e di Internet hanno inaugurato la terza rivoluzione industriale

e posto le basi per la rapida ascesa dell'IA, i Big Data e l'analisi da essi generata ci hanno permesso di portare la nostra intelligenza a nuovi livelli [25]. L'IA è ora considerata una delle principali invenzioni della quarta rivoluzione industriale [26]. Esperimenti che avrebbero richiesto anni per essere eseguiti sono ora fattibili e spesso poco costosi grazie ai recenti progressi nei dati e nella metodologia. Come risultato di queste analisi sperimentali vengono generati dati grezzi in vari formati. La capacità di memorizzare e analizzare i dati con l'aiuto dell'IA ha creato nuove possibilità per la comunità accademica, i ricercatori scientifici e l'industria biotecnologica. Diverse applicazioni dell'IA sono utilizzate in biologia, tra cui l'identificazione precisa della geometria 3D di molecole biologiche come le proteine, uno dei compiti più critici e utili nella ricerca biologica. Inoltre, nella scienza biologica, l'IA svolge un ruolo fondamentale nel promuovere l'innovazione non solo nei laboratori, ma anche nell'intero ciclo di vita di un farmaco o di un prodotto chimico [27]. Inoltre, gli strumenti e le applicazioni basati sull'IA aiutano ad automatizzare complicate procedure di produzione, soddisfacendo così la domanda in rapida crescita di farmaci, prodotti chimici per l'industria, alimenti e altre materie prime di origine biologica. Il ML, un sottoinsieme dell'IA, aiuta a prevedere i risultati eseguendo permutazioni e combinazioni massicce di insiemi di dati disponibili per le molecole di farmaci, al fine di determinare la combinazione migliore senza affidarsi ai tradizionali metodi manuali in laboratorio [28]. Sebbene i metodi tradizionali basati su modelli siano ancora utili per l'analisi dei dati biologici, non hanno la capacità di utilizzare grandi quantità di dati disponibili, o addirittura big data, per scoprire informazioni, prevedere il comportamento dei dati e comprendere collegamenti complessi tra i dati. L'uso estensivo dei big data sta diventando sempre più importante nelle biotecnologie e nella bioinformatica, in quanto i dati continuano a crescere e diventano disponibili per l'analisi da parte di accademici e scienziati di tutto il mondo [29]. Questi dati sono quantificati in termini di multi-omica, come genomica, trascrittomica, proteomica e metabolomica, provenienti da diverse fonti biologiche e devono essere annotati e analizzati in modo

appropriato per comprendere dei sistemi biologici complessi. L'intelligenza artificiale e i progetti di reti neurali profonde potrebbero analizzare in modo efficiente i dati genomici per determinare le basi genetiche di un tratto e scoprire i marcatori genetici legati a determinati tratti [30,31]. L'uso dell'intelligenza artificiale può aiutare a decifrare i legami complessi tra le diverse informazioni nascoste nei dati, per ricavarne spunti significativi. Di conseguenza, l'incorporazione di approcci di IA è ora ampiamente osservata nel campo delle scienze biologiche e si prevede un ulteriore aumento nel prossimo futuro grazie al progredire di questa tecnologia [2]. Inoltre, le immagini mediche e le risposte ai farmaci forniscono dati complessi ma significativi e richiedono programmi algoritmici efficienti per analizzarli. Per questo motivo, l'IA basata su ML e DL sta raccogliendo molta attenzione grazie alle sue capacità di elaborare più velocemente dati enormi ed estrarre informazioni significative. L'elaborazione di immagini digitali, la progettazione di farmaci e i test farmacologici virtuali basati sull'IA potrebbero trasformare la scienza medica nel prossimo futuro [32,33].

L'articolo in esame evidenzia come l'Intelligenza Artificiale e i suoi componenti possano essere utilizzati nei settori medico, agricolo e bioindustriale per rendere la vita umana più sostenibile.

2. L'INTELLIGENZA ARTIFICIALE NELLA SCIENZA MEDICA

I progressi della scienza medica e delle biotecnologie hanno aperto nuove strade per lo sviluppo di farmaci e antibiotici. L'IA ha un enorme potenziale per applicazioni diffuse nell'industria farmaceutica (Figura 4). Con l'IA è possibile scoprire nuove molecole terapeutiche basate su strutture bersaglio note [34]. Un ramo dell'IA noto come ML è comunemente impiegato nella diagnosi delle malattie, poiché sfrutta i risultati dei test diagnostici per migliorare l'accuratezza dei risultati [35]. L'IA consente ai ricercatori di gestire questioni impegnative, tra cui l'epidemiologia quantitativa e predittiva, i farmaci di precisione e le interazioni ospite-patogeno [36]. L'IA può contribuire all'individuazione e alla diagnosi

delle malattie e rendere il codice informatico più accessibile ai non addetti ai lavori [37]. L'epidemiologia predittiva, la medicina di precisione basata sull'individuo e l'analisi delle interazioni ospite-patogeno sono esempi di aree di ricerca che potrebbero trarre vantaggio dai progressi dell'apprendimento automatico e profondo [38]. Questi approcci aiutano a diagnosticare le malattie e a identificare i singoli casi, a fare previsioni più accurate e a ridurre gli errori, a prendere decisioni più rapide e a migliorare l'analisi dei rischi (Figura 4). Il numero crescente di biomarcatori tissutali e la complessità delle loro valutazioni promuovono in modo significativo l'uso di tecniche basate sull'IA. Questi biomarcatori basati sull'IA aiutano i medici nella previsione e nell'analisi della diagnosi, della risposta del paziente al trattamento e della sua sopravvivenza [39]. Grazie alla rappresentazione della conoscenza e alla modellazione del ragionamento, è possibile ottenere modelli più realistici di sistemi socio-biologici complessi [40]. I metodi basati sul ML possono essere utilizzati anche per migliorare l'efficienza e l'affidabilità dei modelli epidemiologici [41,42]. Ad esempio, i progressi del ML hanno aiutato a sviluppare modelli basati su programmi algoritmici di dieci parametri cellulari che possono distinguere accuratamente i tumori benigni da quelli maligni [43].

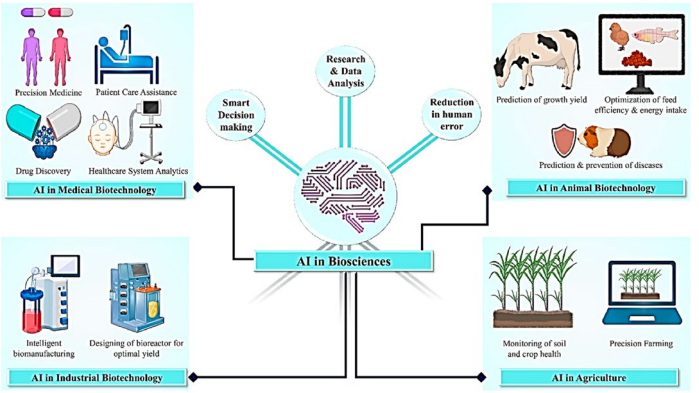


Figura 4. Schema che illustra le possibilità di applicazione dell'intelligenza artificiale nelle discipline della salute, dell'agricoltura, delle scienze animali e delle biotecnologie industriali.

Nella medicina di precisione è importante tenere conto delle differenze individuali in termini di genetica, ambiente e stile di vita [44]. I medici riconoscono che la composizione metabolica, fisica, fisiologica e genetica di un individuo influisce sul modo in cui il suo corpo risponde ai farmaci. Ciononostante, attualmente utilizziamo un approccio generico che tratta tutti i pazienti con lo stesso farmaco, indipendentemente dalle loro diverse condizioni. Tuttavia, grazie soprattutto ai progressi dell'intelligenza artificiale, si sta evolvendo una nuova era di medicina personalizzata, in cui i farmaci vengono adattati alle esigenze e alla capacità di adattamento dell'organismo. Sebbene la transizione sembri semplice, comporta una quantità significativa di raccolta, elaborazione, manutenzione ed esecuzione dei dati [45]. Inoltre, milioni di analisi predittive saranno incluse nel processo per identificare le migliori molecole terapeutiche candidate per un caso specifico. Utilizzando questa strategia, medici e clinici potranno prevedere meglio quali strategie di trattamento e prevenzione delle malattie saranno più efficaci per particolari gruppi di pazienti (Figura 4). I ricercatori potrebbero utilizzare l'IA negli studi sul DNA, sull'RNA e sulle proteine per visualizzare meglio gli effetti delle dosi di farmaco sui tessuti viventi nel tempo e per riorganizzare il monitoraggio durante la terapia [46,47]. Basandosi sull'IA, IBM Watson assiste nella creazione del piano di trattamento appropriato per un paziente in base alla sua storia medica e ai suoi dati personali, compresa la composizione genetica [48]. Un sistema di medicina personalizzata basato sull'IA non solo ridurrà i costi del trattamento, ma minimizzerà anche gli effetti collaterali dei farmaci [49]. Oltre a risparmiare tempo e a migliorare la cura del paziente, l'IA può anche semplificare l'editing genetico, la radiografia e la procedura di pianificazione della gestione dei farmaci [50]. Inoltre, le cartelle cliniche elettroniche (EHR) possono essere migliorate con sistemi basati sull'evidenza che siano di supporto alle decisioni cliniche [44,51,52]. L'IA comporta un'enorme capacità di elaborazione (supercomputer), algoritmi in grado di apprendere a una velocità fenomenale (deep learning) e una nuova strategia che utilizza

le capacità cognitive dei medici (Tabella 1). Questa tecnica può contribuire allo sviluppo di modelli teorici innovativi della fisiopatologia delle malattie e può aiutare a prevedere i principali effetti avversi dei farmaci a lungo termine [53]. In uno studio recente, un approccio basato sull'IA si è rivelato molto utile per l'identificazione, la diagnosi, la prognosi e il trattamento precoce della miopia [54]. In cardiologia, dermatologia e oncologia, gli algoritmi di deep learning superano i medici almeno nella diagnosi delle malattie [55,56,57]. È evidente che gli algoritmi informatici sono in grado di rilevare il cancro al seno metastatico nelle biopsie del linfonodo sentinella in immagini a vetrino completo con un tasso di accuratezza superiore al 91%, che è salito al 99,5% quando sono stati aggiunti gli input del medico [58]. Una delle applicazioni comuni dell'IA nell'analisi del rischio è la diagnosi di malfunzionamento del cuore attraverso l'imaging cardiovascolare. Essa comprende il monitoraggio automatico di eventuali deviazioni dalle condizioni normali basate sull'elaborazione delle immagini, la funzione miocardica e il rilevamento e l'analisi delle placche aterosclerotiche coronariche [59]. È stato utilizzato l'algoritmo YOLOV3 per la segmentazione di immagini mediche basata sull'intelligenza artificiale per la stampa 3D e la visualizzazione 3D a occhio nudo per individuare la prostata nelle immagini di risonanza magnetica pesate in T2 (AIMIS3D) [59]. Esistono diverse variabili che potrebbero essere analizzate in modo efficiente attraverso l'IA, come ad esempio determinare quali patologie sono resistenti a determinati antibiotici e non ad altri [60]. Tali analisi possono supportare i medici e ridurre significativamente i test non necessari e i costi delle cure mediche.

Malattie studiate	Algoritmo	Modalità	Risultati	Riferimenti
AMD	Modello predittivo basato su ML	Dati clinici	Un modello predittivo basato sull'intelligenza artificiale è stato in grado di prevedere la progressione dell'AMD con elevata precisione	[61]

Malattia di Alzheimer	RF, SHAP	Dati clinici e di imaging	Il modello AI è stato in grado di rilevare e prevedere con precisione la progressione della malattia di Alzheimer con un'accuratezza del 93,95% nel primo strato e dell'87,08% nel secondo strato.	[62]
COVID-19	PA	Dati clinici	È stata raggiunta un'accuratezza del 70-80% nel predire i casi di COVID-19 grave.	[63]
Cancro ovarico	ANN	Dati clinici	È stata raggiunta un'accuratezza del 93% nel predire la sopravvivenza delle pazienti affette da carcinoma ovarico e un'accuratezza del 77% nel predire l'esito chirurgico.	[64]
Cancro polmonare	LCP-CNN, modello Brock	Dati clinici	LCP-CNN è stato in grado di prevedere la malignità dei noduli polmonari con una maggiore accuratezza e minori risultati falsi negativi rispetto al modello Brock.	[65]
Influenza	IAT-BPNN	Dati CDC e set di dati Twitter	IAT-BPNN è stato in grado di prevedere la malattia simil-influenzale in una popolazione di grandi dimensioni con un'elevata accuratezza.	[66]

Tabella 1. L'intelligenza artificiale nel rilevamento delle malattie e nella modellazione delle previsioni.

È importante sottolineare l'importanza di combinare questi algoritmi con le competenze mediche (Tabella 1). È possibile scoprire nuovi composti farmaceutici attraverso l'analisi dei dati utilizzando l'IA, il che riduce la necessità di sperimentazioni cliniche, consentendo di immettere più rapidamente sul mercato i farmaci senza comprometterne la sicurezza [32]. Inoltre, con l'aiuto dell'IA potremmo essere in grado di prevedere con notevole anticipo l'insorgenza di malattie geneticamente predisposte [50]. I pazienti potranno anche prevenire e trattare alcune malattie ereditarie.

Una delle applicazioni dell'IA nell'industria farmaceutica è "Open Targets", uno sforzo strategico relativamente nuovo per esplorare la relazione tra i bersagli dei farmaci e le malattie, nonché il modo in cui alcuni geni sono collegati alle malattie [67].

SPIDER è un'altra tecnica di intelligenza artificiale progettata per determinare il ruolo dei prodotti naturali nella scoperta dei farmaci [68]. Inoltre, gli studi di relazione quantitativa struttura-attività (QSAR) sono particolarmente utili per creare nuovi farmaci efficaci in un periodo di tempo molto breve utilizzando uno strumento di simulazione al computer [69]. In un recente studio è stato utilizzato un modello QSAR basato su una rete neurale artificiale (ANN) a base radiale (RBF) addestrata con la tecnica dell'ottimizzazione con sciami di particelle (PSO) per prevedere i valori di pKa di 74 diversi tipi di farmaci [69]. L'elaborazione del linguaggio naturale (NLP), il ML e l'automazione dei processi robotici sono chiaramente le tre aree chiave di avanzamento dell'IA nel campo della medicina [70,71]. L'elaborazione del linguaggio naturale è stata recentemente utilizzata per migliorare l'analisi della colonscopia, migliorando l'individuazione accurata di adenomi e polipi [72]. Inoltre, un approccio ML può essere utilizzato per prevedere malattie come la fibrillazione atriale e le infezioni del tratto urinario in determinati gruppi di pazienti, utilizzando modelli come la macchina a vettori di supporto (SVM) basati sulle caratteristiche cliniche della malattia [73,74,75]. Iniziative simili sono state utilizzate per migliorare la prognosi delle malattie cardiache utilizzando una tecnologia per il rilevamento dei rumori cardiaci [76]. La FDA ha già approvato fino a 29 dispositivi medici e algoritmi basati sull'IA in vari campi delle scienze mediche [77].

Il primo modello basato sull'intelligenza artificiale approvato dalla FDA nel settore sanitario è stato un modello diagnostico basato su un sistema autonomo di intelligenza artificiale, IDx-DR. Questo modello è stato utilizzato con successo per rilevare la retinopatia diabetica con una sensibilità, specificità e capacità di immagine rispettivamente dell'87,2%, del 90,7% e del 96,1%, su un campione di 819 soggetti in 10 unità di cure primarie negli Stati Uniti. Il modello è stato addestrato con un campione di dati diversificato composto da individui di età, etnia e sesso diversi, riducendo così al minimo le possibilità di errore nei vari gruppi [78]. Sono stati condotti anche diversi studi clinici rando-

mizzati (RCT) per testare l'efficacia e la sicurezza dei modelli AI e ML nella pratica clinica. In un RCT (numero di registrazione: ChiCTR-DDD-17012221), è stato valutato l'impatto di un algoritmo di identificazione automatica dei polipi basato sul deepleaning dell'accuratezza del rilevamento dei polipi e sul tasso di rilevamento degli adenomi (ADR). In questo RCT, i pazienti successivi sono stati assegnati in modo casuale alla colonscopia con o senza l'aiuto del modello di identificazione automatica dei polipi che forniva una notifica ottica e un avviso sonoro al momento della scoperta del polipo. I risultati ottenuti dai pazienti sottoposti al sistema di rilevamento automatico basato sull'intelligenza artificiale hanno superato le coorti di controllo per quanto riguarda l'ADR e la quantità media di adenomi e polipi rilevati per coloscopia. Questa tecnologia automatizzata può quindi essere adeguata nei trattamenti e nelle pratiche di routine per una migliore identificazione dei polipi del colon grazie alla sua grande sensibilità, all'elevata precisione e alla stabilità dei risultati [79]. L'introduzione di sistemi di intelligenza artificiale nel processo decisionale medico ha anche permesso di migliorare il rapporto costo-efficacia di un trattamento medico completo. In uno studio, l'uso di un algoritmo decisionale basato sulla procalcitonina (PCTDA) per i pazienti ospedalizzati affetti da sepsi e infezioni delle basse vie respiratorie ha portato a una riduzione della durata della degenza, a una minore somministrazione di antibiotici, a un minor numero di periodi di ventilazione artificiale e a una diminuzione del numero di pazienti con infezioni e resistenza agli antibiotici. In media, il trattamento basato su PCTDA ha comportato una riduzione del 49% e del 23% delle spese complessive rispetto al trattamento convenzionale per sepsi e infezioni delle basse vie respiratorie [80]. L'industria farmaceutica riuscirà a cogliere meglio le informazioni genetiche grazie al miglioramento delle competenze di AI e ML (Figura 4). Evidentemente, se integrata con ML e NLP, l'automazione robotica dei processi ha applicazioni significative e ha il potenziale per rimodellare la scienza medica nel prossimo futuro [81]. Nonostante

gli enormi progressi osservati, c'è ancora molto lavoro da fare prima che la terapia basata sull'IA diventi realtà.

3. L'INTELLIGENZA ARTIFICIALE NELLE BIOTECNOLOGIE AGRICOLE

[...]

4. IA E BIOTECNOLOGIA INDUSTRIALE

La biotecnologia industriale, talvolta nota come biotecnologia bianca, è la moderna applicazione della biotecnologia alla lavorazione e alla produzione sostenibile di prodotti di base, chimici e combustibili da fonti rinnovabili utilizzando cellule vive e i loro enzimi. La domanda di prodotti chimici industriali, medicinali, prodotti chimici per uso alimentare e altre materie prime legate alla biochimica è aumentata notevolmente negli ultimi dieci anni [121]. Le tecnologie basate su ML e AI possono contribuire alla progettazione di nuovi farmaci e all'identificazione della loro efficacia e dei loro effetti avversi prima della loro effettiva produzione, riducendo drasticamente il tempo necessario per portare un farmaco dal laboratorio al mercato per la gente comune [32]. I microrganismi e le cellule vegetali/animali sono utilizzati nei processi biotecnologici per realizzare prodotti in diversi settori, tra cui farmaci, prodotti farmaceutici, alimenti e mangimi, disinfettanti, pasta di legno e tessuti. Per rilevare le interruzioni, ottimizzare i macchinari per una produzione efficiente e migliorare la qualità dei prodotti e l'Internet delle cose potrebbero essere utilizzati in modo efficace il ML e l'IA [122]. I modelli informatici basati sull'IA stanno diventando sempre più diffusi e la robotica e il machine learning potrebbero essere utilizzati per sviluppare le migliori condizioni di crescita ottimali per i ceppi, nonché il grado di ottenimento di prodotti di valore (Figura 4). Ad esempio, approcci basati sull'IA o su metodologie di superficie di risposta (RSM) sono stati utilizzati per la produzione

ad alto livello di amilasi da *Rhizopus microsporous*, utilizzando vari scarti agroindustriali per progettare esperimenti in condizioni ottimali [123]. Allo stesso modo, degli algoritmi di intelligenza artificiale come le reti neurali artificiali (ANN) e gli algoritmi genetici (GA) sono stati integrati per l'ottimizzazione dei terreni di fermentazione per produrre glucansucrasi da *Leuconostoc dextranicum*. Il modello integrato ANN-GA ha visto un aumento del 6% dell'attività della glucansicrasi rispetto a un approccio di previsione basato sulla regressione [124]. Recentemente è stato applicato il modello integrato ANN-GA per l'ottimizzazione della produzione di cellulasi da parte di *Trichoderma stromaticum* in fermentazione allo stato solido; dopo l'ottimizzazione con il modello AI è stato ottenuto un aumento di 31,58 volte della produzione di cellulasi [125].

Le tecnologie basate sull'intelligenza artificiale sono state utilizzate anche per scalare e ottimizzare i bioprocessi per la produzione di enzimi su una scala di prova. In uno studio è stato eseguito un metodo a basso costo per aumentare la sintesi di laccasi extracellulare da Staphylococcus arlettae utilizzando gli scarti del tè. RSM e ANN accoppiati con GA sono stati due metodi statistici consecutivi impiegati per aumentare la produzione di enzimi e hanno portato a un aumento di sedici volte della resa enzimatica. Inoltre, è stato creato un bioprocesso su scala pilota utilizzando i parametri ideali identificati dalla GA, ossia rifiuti di tè (2,5%) NaCl (4,95 mM), L-DOPA (5,65 mM) e temperatura di 37°C, che ha migliorato la produzione di enzimi di 72 volte [126]. Inoltre, alcuni modelli di intelligenza artificiale basati sul sistema esperto fuzzy sono in grado di monitorare gli impianti di trattamento delle acque reflue su scala pilota [127].

Il biocarburante è uno dei più importanti bioprodotti per i quali il processo di produzione industriale può essere migliorato utilizzando il ML e l'AI per ottenere il massimo rendimento. Nel settore delle bioenergie, gli approcci basati sull'IA sono stati utilizzati per prevedere le proprietà delle materie prime della biomassa, gli usi finali della bioenergia e le catene di approvvigionamento della bioenergia e hanno sviluppato un modello integrato

con il metodo ANN-Taguchi per la previsione e la massimizzazione della produzione di biocarburante attraverso la torrefazione e la pirolisi [128,129]. L'ottimizzazione e la progettazione dei fattori sperimentali sono state eseguite utilizzando il metodo Taguchi che ha portato al raggiungimento della massima resa di biocarburante fino al 99,42%, mentre ANN ha mostrato una regressione lineare di 0,9999 per il biochar e di 0,9998 per i bio-oli.

I modelli integrati ANN-GA sono stati utilizzati per la modellazione e l'ottimizzazione del processo di metanolisi dei gusci di arachidi di scarto per la produzione di biocarburanti. La resa in biocarburante ottimizzata dal modello RSM è stata del 16,49%, mentre quella del modello ANN-GA è stata del 17,61%. Ciò dimostra che il modello integrato ANN-GA ha un potenziale di ottimizzazione migliore rispetto al modello RSM da solo [130]. Modelli di bioprocesso basati su ML sono stati costruiti anche con l'aiuto di metodi basati sull'AI come ANN, CNN, (reti di memoria a breve termine) LSTM, kNN (k-nearest neighbors) e RF (random forests) per prevedere l'accumulo di carboidrati nella biomassa di cianobatteri coltivati in acque reflue per la produzione di biocarburanti. I risultati migliori per l'approssimazione della dinamica del sistema sono stati ottenuti con una 1D-CNN con un errore quadratico medio di 0,0028 [131]. I tessuti, i nuovi prodotti chimici e la sintesi di biopolimeri biodegradabili potrebbero beneficiare di processi simili [132]. Inoltre, l'intelligenza artificiale può essere utilizzata per contribuire allo sviluppo di tecniche di sintesi per questi prodotti biochimici che producono la massima resa con la minima quantità di input (Figura 4). Inoltre, l'IA potrebbe aiutare a prevedere in tempo reale la domanda di mercato di farmaci o sostanze chimiche. L'IA e il ML hanno contribuito anche alla produzione di metaboliti. L'ingegneria metabolica dei sistemi è un processo che aiuta a produrre rapidamente ceppi microbici ad alte prestazioni per la produzione a lungo termine di sostanze chimiche e minerali. La crescente disponibilità di big data biologici, come i dati omici, ha portato all'applicazione di tecniche di ML in varie fasi dell'ingegneria metabolica dei sistemi, come la selezione del cep-

po ospite, la ricostruzione della via metabolica, l'ottimizzazione del flusso metabolico e la fermentazione [19]. Diversi algoritmi di apprendimento automatico, tra cui il deep learning, hanno facilitato l'ottimizzazione dei parametri dei bioprocessi e l'esplorazione di un più ampio spazio metabolico legato alla biosintesi di un bioprodotto target [133]. Questa tendenza sta spingendo anche le aziende biotecnologiche ad adottare tecniche di ML nella creazione dei loro sistemi di produzione e delle loro tecnologie di piattaforma [134]. Nell'industria birraria, l'IA ha dimostrato un potenziale promettente per superare le carenze del settore e migliorare la produzione attraverso l'accumulo di conoscenze e il controllo automatizzato. In uno studio, sono stati costruiti modelli di intelligenza artificiale utilizzando profili aromatici e dati spettroscopici ottenuti da alcolici commerciali per valutare i tratti qualitativi e l'aroma della birra. I modelli intelligenti hanno dato luogo a previsioni altamente accurate per sei aromi di birra [135]. Sono state sviluppate anche tecnologie di e-nose intelligente basate su modelli ANN per valutare la presenza di diverse sostanze chimiche come etanolo, metano, monossido di carbonio, idrogeno solforato, ammoniaca e così via nella birra [136]. Uno studio si è occupato dello sviluppo di un programma informatico che simulava il funzionamento di una rete neurale di percezione feed-forward a tre strati, altamente personalizzabile che, utilizzando i dati di esperimenti precedenti, era in grado di prevedere le variazioni dei parametri della fermentazione alcolica del vino bianco. Questo lavoro ha fornito un approccio adeguato per la digitalizzazione dei processi di produzione della birra, consentendo così la coesistenza con altre strutture intelligenti e basate sulla conoscenza [137]. Un altro studio ha portato allo sviluppo di un approccio innovativo basato sulla conoscenza per il controllo della fermentazione in batch dell'alcol impiegato nella produzione di vino bianco. Le fonti primarie di informazioni utilizzate per lo sviluppo del modello di intelligenza artificiale sono stati diversi casi di studio e risultati sperimentali, nonché le conoscenze ottenute dagli esperti del birrificio in merito a diversi parametri per l'ottimizzazione e il controllo del processo com-

plessivo. Utilizzando il software di monitoraggio, regolazione e acquisizione dati del bioreattore di fermentazione, è stata sviluppata un'applicazione per il controllo automatizzato del processo [138]. L'ulteriore integrazione di sistemi di controllo, processi e progressi innovativi può essere notevolmente facilitata da questi tipi di modelli di intelligenza artificiale, in un'ottica di sviluppo sostenibile.

5. SFIDE E LIMITI

Nonostante il loro immenso potenziale, le tecnologie basate sull'IA devono ancora farsi strada nella pratica quotidiana. I modelli di IA possono migliorare l'accessibilità di vari settori biologici, ma possono anche esacerbare le differenze preesistenti. Poiché i modelli di IA sono estremamente dipendenti dai set di dati su cui sono sviluppati e dalle etichette ad essi collegate, potrebbero essere rafforzate le distorsioni sui soggetti sottorappresentati negli algoritmi di apprendimento [139]. Per valutare correttamente la resilienza di alcune reti neurali profonde è necessario considerare diversi fattori. Per lo sviluppo di modelli di intelligenza artificiale è necessario creare, recuperare e pulire i metadati. I programmi dovrebbero inoltre essere progettati e valutati sotto la supervisione di professionisti del settore per l'analisi e la correzione delle imprecisioni commesse nella pratica [140]. Nonostante i notevoli progressi compiuti negli ultimi anni nella progettazione di modelli basati sull'IA e sul ML, pochi sono stati incorporati nell'assistenza sanitaria e molte prospettive di adozione di questi modelli per l'uso quotidiano rimangono inesplorate. Ad esempio le CNN all'inizio, dal 2015, erano usate perlopiù in progetti di studio sulle radiografie dentali, mentre i primi usi clinici di questi strumenti sono emersi solo di recente [141]. L'indisponibilità e l'inaccessibilità dei dati clinici, dovuta alle policy organizzative, la scarsa riproducibilità dell'elaborazione dei set di dati e della valutazione degli esiti e le preoccupazioni sulla responsabilità e trasparenza verso i pazienti rimangono gli

ostacoli più comuni all'adattamento dell'IA alle pratiche mediche e odontoiatriche di routine [142]. Inoltre, è stato riferito che diversi modelli non sono accurati nel predire la diagnosi clinica. Ad esempio, è stato sviluppato un algoritmo di IA in grado di diagnosticare e classificare le radiografie del torace utilizzando l'NLP per le registrazioni radiologiche [82,143]. Queste classificazioni sono state successivamente usate nell'addestramento di una rete di apprendimento profondo per rilevare anomalie nelle immagini, concentrandosi sul riconoscimento di uno pneumotorace [144]. Tuttavia, dopo un esame approfondito, la presenza di un tubo toracico nella maggior parte dei referti identificati come pneumotorace ha sollevato il dubbio che l'algoritmo abbia riconosciuto dei tubi toracici anziché uno pneumotorace come previsto [143]. Un altro esempio di risultati non interpretativi di un sistema clinico basato sull'IA è DeepGestalt, uno strumento per l'analisi della dismorfologia facciale. Questo strumento ha ottenuto scarsi risultati nell'identificazione delle persone con sindrome di Down di origine africana (36,8%) rispetto a quelle di origine europea (80%) [145]. Le diagnosi di sindrome di Down tra le persone di origine africana sono aumentate al 94,7% quando il modello è stato riqualificato utilizzando casi di persone affette dalla sindrome [145]. A causa di varie marginalità nei set di dati di addestramento, la modellazione della suscettibilità alle malattie genetiche è anche predisposta a prestazioni differenziali tra gruppi demografici [146]. Inoltre, è stato osservato che mentre gli approcci ML possono dare risultati migliori negli studi per lo sviluppo di modelli di previsione del rischio di malattia, la presentazione dei dati può essere più complessa. È anche possibile che il tempo di calcolo richiesto dagli approcci ML vari a seconda della dimensione dei dati [147]. Pertanto, è fondamentale riconoscere che l'utilizzo di approcci basati sull'IA non porterà sempre a una categorizzazione migliore o a una previsione migliore rispetto ai metodi attuali. L'IA è uno strumento che deve essere impiegato nel contesto appropriato per affrontare una domanda pertinente o risolvere un problema significativo [148]. Allo stesso modo, in altri campi biologici come l'agricol-

tura, l'automazione delle pratiche che impiegano approcci basati sull'IA e sul ML sfrutta un grande potenziale per un'agricoltura sostenibile. Tuttavia, nel settore agricolo, la raccolta, l'analisi e l'utilizzo dei dati per la produttività presentano una serie di ostacoli. La privacy e la sicurezza dei dati sono le due sfide principali che gli agricoltori devono affrontare per sopravvivere nell'era digitale. Nella maggior parte dei casi, gli agricoltori non sono informati della raccolta, dell'utilizzo e, cosa più preoccupante, degli scopi per cui i loro dati personali vengono utilizzati [149]. Il data mining consente alle aziende di affidarsi agli individui per acquisire dati di massa sull'agricoltura, che possono essere sufficienti per sviluppare e valutare i quadri comportamentali e psichiatrici degli intervistati [150]. Per evitare che i dati vengano utilizzati in modo improprio, gli agricoltori devono avere la certezza che le loro informazioni saranno utilizzate per generare idee innovative e soluzioni agricole piuttosto che per ottenere un vantaggio competitivo. Come menzionato altrove, la tecnologia dei droni basata sull'intelligenza artificiale è emersa come un approccio molto efficace in agricoltura [87]. Tuttavia, i droni, in particolare quelli dotati di obiettivi ad alta risoluzione, telecamere a infrarossi, programmi competenti e sensori, sono molto costosi per i piccoli agricoltori. Inoltre, per utilizzare i droni è necessaria un'autorizzazione in base alle disposizioni operative e normative del diritto fondiario [151]. Inoltre, le condizioni meteorologiche influenzano enormemente il funzionamento dei droni [152]. Le metodologie tradizionali di data mining sono state sviluppate principalmente per gli insiemi di dati relazionali; tuttavia, non sono completamente adeguate per i dati geograficamente sparsi [153]. Per rivoluzionare l'agricoltura con le tecnologie basate sull'IA, sono necessari degli approcci innovativi di data mining.

Nel settore delle biotecnologie industriali, la definizione di protocolli definiti e praticabili per l'adozione di un algoritmo e la valutazione delle dimensioni del set di dati rimane una sfida importante. Per progettare tali protocolli, sarebbe necessario avere una conoscenza approfondita degli effetti e dell'efficacia dei vari algoritmi e dei set di dati di addestramento per affrontare le

numerose sfide della bioindustria. Inoltre, sono ancora necessari una maggiore accessibilità, una buona documentazione e dei metodi di acquisizione dei dati di qualità superiore per sviluppare, far funzionare e ottimizzare i sistemi bioenergetici e i progetti di bioreattori [128]. In alcuni modelli di IA, quando l'input è inadeguato, in particolare per insiemi di dati di grandi dimensioni, l'algoritmo può richiamare ogni singola variabile come istanza speciale invece di apprendere le informazioni, con conseguenti errori e calo dell'efficacia dell'addestramento [154]. Inoltre, numerosi sistemi rappresentati da RNA sono spesso criticati perché appaiono come delle scatole nere. Tuttavia, la scarsità di lavori comparativi tra i diversi progetti di AI-ML rende difficile presentare una direzione chiara per gli studi futuri o per l'applicazione pratica [155]. Esistono ancora sfide che devono essere superate, tra cui l'inefficiente integrazione dei dati, dovuta alla diversità degli insiemi di dati, compresi i dati candidati, i metadati, i dati elaborati e i dati grezzi, e alla mancanza di competenze adeguate e di esperienza in materia [156]. In questo contesto, è necessario superare queste ambiguità utilizzando nuovi algoritmi di intelligenza artificiale per allineare i risultati previsti agli studi empirici [157]. Insomma, per sviluppare modelli basati su IA e ML per il monitoraggio e il controllo in tempo reale di bioreattori e bioprocessi sono necessari set di dati più ampi e studi relativi.

6. CONCLUSIONI

Una delle grandi conquiste che abbiamo visto nell'era dell'Industria 4.0 è la capacità di una macchina di replicare le capacità dei sistemi viventi, in particolare l'intelligenza di un essere umano. La capacità di riconoscere gli oggetti e prendere decisioni è una caratteristica fondamentale dei sistemi biologici. Attualmente l'IA è in grado di riconoscere gli oggetti e prendere decisioni utilizzando molte delle capacità cognitive e percettive dei sistemi viventi. Il potenziale dell'IA potrebbe essere utilizzato per il mondo biologico, compresa la ricerca medica, l'agricoltura e le

bioindustrie, per il nostro stile di vita sostenibile. La previsione e l'identificazione precoce delle malattie e il loro trattamento preciso basato sulla medicina personalizzata, anche quando le malattie sono in condizioni asintomatiche, sono esempi di aree chiave della scienza medica che potrebbero beneficiare dell'IA. Questo non solo salverebbe milioni di vite, ma ridurrebbe anche i costi medici. Oltre al campo medico, di recente sono stati sviluppati algoritmi e programmi efficienti basati sull'IA per garantire input e output efficaci in agricoltura, una pratica nota come agricoltura di precisione. Potrebbero essere rivoluzionate dall'IA anche alcune pratiche agricole come la gestione del suolo, l'analisi del fabbisogno idrico, la selezione dell'esatto fabbisogno di fertilizzanti, pesticidi, insetticidi, erbicidi, i pronostici di resa e la gestione complessiva delle colture. Ciò contribuirebbe a soddisfare la crescente domanda di cibo della popolazione mondiale. Quando si parla di produzione su larga scala, molti fattori variabili portano all'aumento dei costi, che rappresentano una sfida importante. Recentemente, i programmi e i modelli informatici basati sull'intelligenza artificiale si sono dimostrati molto efficienti nell'ottimizzare le condizioni adatte per ottenere il massimo prodotto desiderato al minimo costo, sia per uso agricolo, medico, biotecnologico o per lo stile di vita. La produzione efficiente di bio-enzimi è solo uno di questi successi ed è facile prevedere come l'industria biotecnologica sarà trasformata dall'applicazione dell'IA, che contribuirà a ridurre i costi di produzione, una delle maggiori sfide che il settore deve affrontare oggi.

4 - SISTEMI BIOMEDICI ASSISTITI DALL'INTELLIGENZA ARTIFICIALE (AI) E DALL'INTERNET DELLE COSE MEDICHE (IOMT) PER UNA SANITÀ INTELLIGENTE

Tratto e tradotto da

Pandiaraj Manickam, Siva Ananth Mariappan, Sindhu Monica Murugesan, Shekhar Hansda, Ajeet Kaushik, Ravikumar Shinde e S. P. Thipperudraswamy, *Artificial Intelligence (AI) and Internet of Medical Things (IoMT) Assisted Biomedical Systems for Intelligent Healthcare, Biosensors*, 2022, *12*(8), 562.

https://doi.org/10.3390/bios12080562

Le parti omesse dal curatore rispetto all'originale sono indicate dal segno **[...]**

ABSTRACT

L'intelligenza artificiale (AI) è una disciplina moderna basata sull'informatica che sviluppa programmi e algoritmi per rendere i dispositivi intelligenti ed efficienti nell'esecuzione di compiti che di solito richiedono un'intelligenza umana qualificata. L'IA comprende vari sottoinsiemi, tra cui l'apprendimento automatico (ML), l'apprendimento profondo (DL), le reti neurali convenzionali, la logica fuzzy e il riconoscimento vocale, con capacità e funzionalità uniche che possono migliorare le prestazioni delle moderne scienze mediche. Questi sistemi intelligenti semplificano l'intervento umano nella diagnosi clinica, nell'imaging medico e nella capacità decisionale. Nella stessa epoca, l'Internet of Medical Things (IoMT) emerge come strumento bioanalitico di nuova generazione che combina dispositivi biomedici collega-

ti in rete con un'applicazione software per migliorare la salute umana. In questa rassegna, discutiamo l'importanza dell'IA nel migliorare le capacità dell'IoMT e dei dispositivi point-of-care (POC) utilizzati in settori sanitari avanzati come la misurazione cardiaca, la diagnosi del cancro e la gestione del diabete. In questo articolo si discute anche del ruolo dell'IA nel supportare gli interventi robotici avanzati sviluppati per applicazioni biomediche avanzate. La posizione e l'importanza dell'IA nel migliorare la funzionalità, l'accuratezza del rilevamento, la capacità decisionale dei dispositivi IoMT e la valutazione dei rischi associati sono qui discusse in modo attento e critico. Questa rassegna comprende anche le sfide tecnologiche e ingegneristiche e le prospettive per i dispositivi IoMT personalizzati integrati nel cloud basati sull'intelligenza artificiale per la progettazione di sistemi biomedici POC efficienti e adatti alla sanità intelligente di prossima generazione.

1. INTRODUZIONE

L'Internet of Medical Things (IoMT) è il sottoinsieme delle tecnologie Internet of Things (IoT) che consiste in dispositivi medici interconnessi in rete per il monitoraggio dell'assistenza sanitaria. I dispositivi IoMT, definiti anche IoT dell'assistenza sanitaria, consentono il monitoraggio dell'assistenza sanitaria senza intervento umano integrando automazione, sensori di interfaccia e intelligenza artificiale basata sull'apprendimento automatico. La tecnologia IoMT collega i pazienti con i medici attraverso i dispositivi medici, consentendo l'accesso remoto per la raccolta, l'elaborazione e la trasmissione di dati medici attraverso una rete protetta. Le tecnologie IoMT contribuiscono a ridurre le degenze ospedaliere non necessarie e quindi i costi sanitari associati, facilitando il monitoraggio wireless dei parametri sanitari. Il segmento delle tecnologie mediche IoMT comprende i dispositivi di monitoraggio personale della salute in tempo reale indossabili e domiciliari e i dispositivi point-of-care (POC) ospe-

dalieri o clinici [1]. La categoria dei dispositivi indossabili per il monitoraggio della salute personale comprende braccialetti intelligenti, tessuti e indumenti elettronici, dispositivi integrati negli smartphone e orologi sportivi per il monitoraggio del fitness e dell'attività fisica [2], come illustrato nella Figura 1.

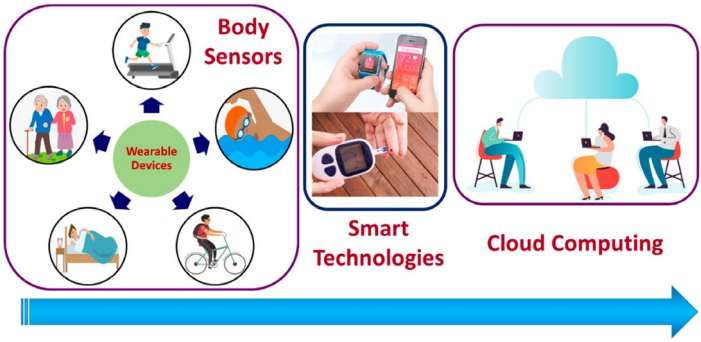


Figura 1. Rappresentazione schematica dei dispositivi IoMT e del trasferimento dei dati nel cloud. I sensori corporei sono quelli fissati direttamente al corpo, incorporati in un tessuto o impiantati nel corpo umano. La tecnologia di rilevamento intelligente viene utilizzata per analizzare i dati raccolti e trasferirli al cloud. Il cloud mette in comunicazione i sensori corporei e il destinatario dei risultati.

Il monitoraggio POC abilitato dall'IoMT per applicazioni cliniche comprende un esame medico su richiesta e il sistema di tele-visita "TyroPro" (https://www.tytocare.com/professionals/, valutato il 22 luglio 2022). I dispositivi indossabili IoMT possono anche monitorare le cadute degli anziani. Le cadute nella popolazione anziana sono inevitabili, ma le loro circostanze possono essere monitorate e prevenute per evitare lesioni croniche. Le tecnologie IoMT contribuiscono a ridurre i ricoveri ospedalieri non necessari e quindi i costi sanitari associati, facilitando il monitoraggio in remoto dei parametri sanitari.

Il monitoraggio dell'assistenza sanitaria tradizionale è in fase di trasformazione e l'assistenza sanitaria digitale consente di mettere nelle mani dei consumatori strumenti di rilevamento automatizzati e prodotti cloud integrati. Questa trasformazione digitale consente ai pazienti, ai medici e alle popolazioni delle

comunità rurali di accedere a servizi sanitari di qualità per ottenere risultati migliori. I dispositivi POC, come gli ultrasuoni, i termometri, i glucometri e i lettori ECG, sono dotati di connettività Internet e di strutture di archiviazione cloud che consentono agli utenti di monitorare la propria salute. I miglioramenti di queste tecnologie sono fondamentali per migliorare l'assistenza sanitaria, regolando le dosi di insulina e collegando i pazienti direttamente con i medici. I centri sanitari avanzati hanno iniziato a utilizzare il concetto di letto intelligente, che può modificare l'angolo e la posizione del letto monitorando la postura del paziente durante il sonno. I dispositivi abilitati all'IoMT contribuiscono anche a trasformare i tradizionali servizi di assistenza sanitaria a domicilio. Ad esempio, il sistema intelligente di distribuzione di farmaci a domicilio carica automaticamente su cloud le informazioni sulla storia medica del paziente. Avverte i medici e i pazienti dei farmaci da assumere e avvisa il medico quando il paziente non prende le medicine. Il progresso tecnologico, l'adattamento industriale e l'urbanizzazione stanno aumentando le richieste del sistema sanitario con l'aumento della popolazione. Il miglioramento dei dispositivi IoMT integrati con smartphone, sensori e attuatori può garantire il monitoraggio periodico dell'assistenza sanitaria: tutti questi aspetti saranno oggetto di questa rassegna.

2. IL RUOLO DELL'INTELLIGENZA ARTIFICIALE NELLA CREAZIONE DI UNA RETE DI SENSORI INTELLIGENTI

L'intelligenza artificiale (IA) è stata oggetto di interesse da parte dei ricercatori e delle industrie biomediche per la sua capacità di elaborare grandi quantità di dati, produrre risultati accurati e controllare i processi per generare un risultato ottimale. L'IA non è una novità, poiché le macchine vengono utilizzate per prendere decisioni e prevedere gli effetti previsti delle malattie e le conseguenze a lungo termine. In questo mondo moderno, la maggior parte delle attività quotidiane è assistita da macchine e algoritmi.

Diversi fattori, come l'equità, la spiegabilità, la responsabilità, l'affidabilità e il consenso, vengono presi in considerazione utilizzando risultati affidabili coordinati da macchine e algoritmi [3]. L'IA può essere interpretata come la capacità del computer o del robot di riprodurre l'intelligenza umana sotto forma di software e algoritmi. L'IA può eseguire processi intellettuali come il ragionamento logico, l'apprendimento basato sulla conoscenza, la scoperta di farmaci, la chirurgia guidata e la diagnostica per immagini avanzata. L'interesse emerso di recente per l'IA può essere nuovo, ma questo concetto era già stato stabilito alla fine degli anni '40 da diversi ricercatori [4,5,6], e le loro idee sono ancora valide come base per le recenti ricerche e invenzioni basate sull'IA. Nel 1972, alcuni ricercatori giapponesi hanno costruito il primo robot umanoide al mondo, "WABOT-1", in grado di comunicare con una persona in giapponese e di misurare distanze e direzioni. Tuttavia, a causa della potenza di calcolo e dei limiti di finanziamento, la ricerca sull'IA è rimasta in un limbo fino alla fine degli anni Novanta. Alla fine degli anni '90, alcune grandi aziende tecnologiche come IBM hanno iniziato a lavorare su modelli basati sull'IA. A metà degli anni 2000, le piattaforme di social network, i fornitori di servizi di posta elettronica, i motori di ricerca e molte altre aziende che elaborano una grande quantità di dati hanno beneficiato di modelli e programmi di IA. Una delle ragioni della recente espansione dell'IA è il miglioramento della potenza di calcolo delle CPU e l'applicabilità delle GPU nel campo dei calcoli. L'altra ragione per l'adozione di un sistema basato sull'IA sono i grandi dati creati dalla domanda degli utenti, necessari per una migliore analisi.

L'apprendimento automatico (ML) è il metodo di intelligenza artificiale più utilizzato per fare previsioni da modelli (Figura 2A). In base alla struttura dell'algoritmo e al metodo di apprendimento, il ML può essere ulteriormente classificato in vari tipi (Figura 2B). I metodi di apprendimento possono essere ulteriormente classificati come apprendimento supervisionato, non supervisionato e per rinforzo. Nell'apprendimento supervisionato, l'algoritmo viene addestrato con i dati di input. L'apprendimento

supervisionato è utilizzato in applicazioni in cui sono disponibili dati storici che possono essere utilizzati per prevedere possibili eventi futuri. Poiché questi algoritmi utilizzano dati storici per l'addestramento, i metodi sono più semplici e accurati. Questi algoritmi possono essere ulteriormente suddivisi in algoritmi di regressione e di classificazione. Gli algoritmi di regressione possono essere utilizzati quando la variabile di ingresso e quella di uscita hanno una relazione, come ad esempio nelle previsioni meteorologiche. Negli algoritmi di classificazione, le variabili di output possono essere classificate in classi come sì-no e vero-falso rispetto alle variabili di input. Grazie a queste caratteristiche, l'apprendimento supervisionato può risolvere un problema reale e prevedere l'output sulla base dei dati disponibili. I metodi di apprendimento non supervisionato possono identificare un modello in ogni set di dati anche se i dati non sono classificati o etichettati correttamente.

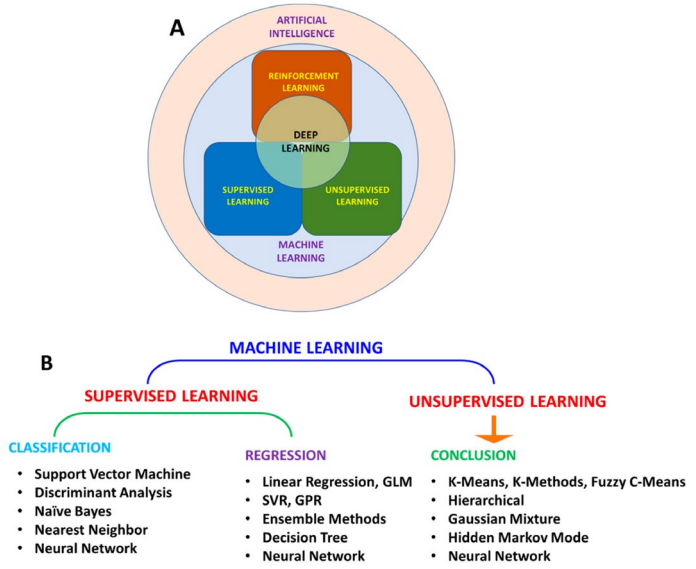


Figura 2. Rappresentazione schematica della relazione tra AI, ML e DL (A); classificazione dell'algoritmo ML (B).

Così facendo gli algoritmi diventano computazionalmente complessi e l'accuratezza diminuisce. Questi algoritmi possono dividere i dati in gruppi in base alle somiglianze e alle differenze dei dati. Possono essere suddivisi in due categorie: (a) clustering e (b) associazione. Nel gruppo degli algoritmi di clustering, i dati vengono raggruppati in cluster basati sulle somiglianze, come ad esempio il comportamento di acquisto di un gruppo di clienti, il che aiuta a commercializzare i prodotti. Un algoritmo di associazione è un algoritmo di apprendimento di regole che trova relazioni tra le variabili in ogni set di dati, ad esempio i modelli di acquisto dei singoli clienti per suggerire prodotti ai clienti. L'apprendimento rinforzato è un metodo di apprendimento basato su premi/penalità. Gli algoritmi assegnano valori positivi ai risultati desiderati e valori negativi agli effetti indesiderati. Questi algoritmi sono difficili da addestrare e richiedono molto tempo. L'apprendimento profondo (DL) è un tipo particolare di ML che insegna ai computer a imitare il comportamento umano. Il DL utilizza le reti neurali (NN), che richiedono molta potenza di calcolo per un problema complesso. Tuttavia, i recenti progressi in termini di potenza di calcolo e di analisi dei dati hanno permesso agli algoritmi DL di osservare, imparare e reagire a situazioni difficili. L'algoritmo DL può adottare approcci di apprendimento supervisionato, non supervisionato o rinforzato in base all'applicazione desiderata. Il miglior esempio di applicazione è rappresentato dai provider di posta elettronica, dove l'intelligenza artificiale viene utilizzata per separare lo spam dalle mail importanti; con ogni nuovo dato, la sua precisione migliora. La maggior parte delle attuali previsioni meteorologiche si basa su modelli di intelligenza artificiale.

I modelli basati sull'IA hanno dimostrato il loro valore nelle industrie farmaceutiche e sanitarie, migliorando l'efficienza nella produzione di farmaci terapeutici, nel monitoraggio della salute in tempo reale e nelle previsioni predittive. L'IA si è già dimostrata promettente nella scoperta dei farmaci e viene implementata in diverse fasi, dalla progettazione allo screening

dei farmaci [7,8,9,10,11]. Nel 2020, il modello DL "Alphafold" ha risolto un problema vecchio di 50 anni prevedendo con precisione la struttura di una proteina a partire dalla sua sequenza di amminoacidi [9,10,11]. Alphafold ha ottenuto risultati migliori, con punteggi TM di 0,7 e superiori per 24 dei 43 domini di modellazione libera, rispetto al secondo miglior metodo di previsione della struttura proteica, che ha raggiunto tale accuratezza solo per 14 domini su 43 in una valutazione cieca [10]. L'IA ha dimostrato di essere uno strumento potenziale per l'individuazione precoce del morbo di Alzheimer [12], del cancro [13], del diabete [14] e delle malattie cardiache, anche in fasi asintomatiche. Esistono applicazioni per le quali l'IA è già utilizzata dalle industrie sanitarie. L'applicazione più diffusa dell'IA nel settore sanitario è la gestione delle cartelle cliniche e della storia dei pazienti. Questi dati possono essere utilizzati dalle app di consultazione digitale, come Babylon nel Regno Unito e Buoy Health negli Stati Uniti, che richiedono un elenco di sintomi, dati sulla storia del paziente e conoscenze mediche comuni per effettuare una diagnosi e consigliare un trattamento.

Jiang et al. hanno pubblicato un elenco dettagliato degli usi di vari algoritmi e metodi di apprendimento dell'IA nella letteratura medica [15]. Secondo questa rassegna, l'apprendimento supervisionato è il metodo di apprendimento consigliato per l'assistenza sanitaria, in quanto fornisce risultati più coerenti dal punto di vista clinico. Jiang et al. hanno inoltre riportato che le macchine a vettori di supporto (SVM) e le reti neurali (NN) sono gli algoritmi basati sull'IA principalmente utilizzati per le applicazioni mediche (Figura 3). Nella Tabella 1 si riporta un confronto generale delle applicazioni, insieme ai vantaggi e agli svantaggi di SVM, NN e altri algoritmi di IA comunemente utilizzati nelle applicazioni biomediche. L'elaborazione del linguaggio naturale (NLP) è un altro campo dell'IA necessario per la piena applicazione dell'IA alla vita reale. L'NLP consente alle macchine/computer di comprendere, analizzare, manipolare e potenzialmente produrre un linguaggio umano. L'input è costi-

tuito da testo scritto o parlato. L'accoppiamento di algoritmi NLP e ML può consentire di svolgere compiti complessi. Esempi comuni di questa categoria sono gli assistenti virtuali come Google Assist, Siri e Alexa. L'NLP viene utilizzato anche per la codifica automatica di documenti clinici [16]. Recentemente, durante il COVID-19, i metodi NLP sono stati messi in pratica per elaborare le note cliniche in un formato leggibile dalla macchina, che aiuta a evidenziare le patologie del paziente e la sua storia medica, oltre ai risultati della valutazione soggettiva, e elabora dei consigli [17].

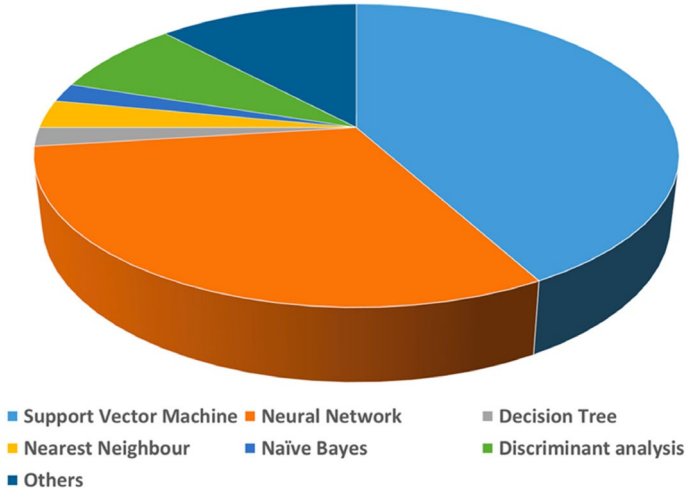


Figura 3. Uso di vari metodi di IA nelle applicazioni mediche.

Algoritmi di intelligenza artificiale	Applicazioni nelle scienze mediche	Vantaggi	Svantaggi
Macchina vettoriale di supporto (SVM)	Imaging dei biomarcatori nei disturbi neurologici e psichiatrici [18] Interfaccia uomo-macchina [19] Diagnosi del cancro [20] Diagnosi precoce della malattia di Alzheimer [21] Monitoraggio cardiaco [22] Previsione dell'infezione del sito chirurgico [23] Monitoraggio del glucosio [24] Chirurgia [25,26,27] Gestione delle risorse per le pandemie [28] Sistema di monitoraggio sanitario [29]	Alta precisione, convergenza più rapida verso la soluzione di un problema, risoluzione di problemi complessi, buona scalabilità per i dati ad alta dimensionalità e necessità di pochi campioni di addestramento.	La selezione di una funzione kernel appropriata è importante, richiede tempi di addestramento più lunghi per insiemi di dati di grandi dimensioni e un elevato costo computazionale. Difficoltà di comprensione e interpretazione del modello finale, dei pesi delle variabili e degli impatti individuali. Problemi nella gestione dei valori mancanti e tendenza all'overfitting.
Rete neurale (NN)	Diagnosi di cancro [13,30,31,32] Identificazione della malattia di Parkinson [33] Monitoraggio cardiaco per immagini [22] Malattia di Alzheimer [34,35] Chirurgia [25,26,27] Applicazioni dei sensori [36,37] Previsione del diabete [38] Interfaccia uomo-macchina [39] Gestione delle risorse per le pandemie [40] Visione artificiale [41]	Algoritmo efficiente, veloce e flessibile. Calcola l'output senza regole programmate, apprende e migliora continuamente. È multitasking e ha ampie applicazioni. Può lavorare con database complessi e non lineari.	Sono necessari tempi di addestramento più lunghi e set di dati di grandi dimensioni. Costo elevato dell'hardware e necessità di programmi lunghi e complessi. L'interpretazione e la modifica sono difficili a causa dell'effetto scatola nera. Incline all'overfitting. L'elevata dipendenza dai dati può dare risultati errati.

Baie naïve (NB)	Previsione della malattia [42] Diagnosi medica [43,44] Gestione delle prestazioni dei sistemi [44] Gestione delle risorse per le pandemie [29]	Facile implementazione, elevata velocità di apprendimento e classificazione. In grado di gestire overfitting, dati rumorosi e valori mancanti. In grado di prevedere la classe di un set di dati di prova. Utile per risolvere problemi di predizione multiclasse.	È distorto per un set di allenamento non ideale. Presenta delle sfide nell'esecuzione di regressioni e caratteristiche co-dipendenti. Non è adatto a problemi complessi.
Vicini di casa K (KNN)	Monitoraggio del glucosio per il diabete [24] Gestione delle risorse per le pandemie [28] Previsione della malattia [45] Diagnosi assistita dal computer [46] Previsione di malattie cardiache [47] Sistema di monitoraggio sanitario [29]	Algoritmo semplice. Nessuna assunzione per le caratteristiche e l'output del set di dati. Efficace contro i dati rumorosi, per la gestione di dati di grandi dimensioni. Prestazioni stabili, elevata velocità di apprendimento e buona gestione dell'overfitting.	Costoso in termini di tempo, sensibile ai dati locali. Accuratezza moderata, velocità di classificazione ridotta. Scarsa gestione dei dati correlati
Albero decisionale (DT)	Monitoraggio del glucosio per il diabete [24] Chirurgia [26,27] Diagnosi medica [44] Gestione delle prestazioni dei sistemi [44] Sistema di monitoraggio sanitario [29]	Molto veloce, efficiente e semplice da capire e interpretare. Può gestire una grande varietà di tipi di dati. Elevata velocità di calcolo, apprendimento e classificazione.	Calcoli complessi. Costoso in termini di tempo e di calcolo. Scarsa capacità di gestire overfitting, dati rumorosi e correlati. Non è in grado di eseguire regressioni e ha un'accuratezza media.

Random Forest (RF)	Previsione della malattia [48,49] Sistema di monitoraggio sanitario [29] Previsione di malattie cardiache [22]	Ottimo per la gestione di dati rumorosi. Alta velocità di classificazione. Ottimo per gestire database grandi ed eterogenei. Definizione automatica delle caratteristiche. La normalizzazione delle caratteristiche in ingresso non è necessaria.	Funzione lavorativa complessa, difficoltà di implementazione. Accuratezza moderata, bassa velocità di apprendimento, scarsa gestione dei valori mancanti. Incline all'overfitting. È importante definire correttamente la profondità e il numero di alberi.
Regressione logistica	Monitoraggio cardiaco per immagini [22] Monitoraggio del glucosio per il diabete [24] Gestione delle risorse per le pandemie [28] Sistema di monitoraggio sanitario [29,50]	Semplicità di implementazione e interpretazione. Buona efficienza formativa. I risultati sono ben calibrati e classificati. Non è necessaria la regolazione empirica dei parametri. Buona precisione per dei set di dati semplici.	Non riesce a risolvere problemi non lineari. Presuppone la linearità delle variabili dipendenti e indipendenti. Incline all'overfitting per insiemi di dati ad alta dimensionalità. Dipende fortemente dai parametri e dalle caratteristiche.

I sistemi basati sull'intelligenza artificiale stanno anche personalizzando i percorsi terapeutici rendendoli più efficaci, trovando farmaci studiati per l'individuo sulla base delle cartelle cliniche e della storia del paziente. I dispositivi indossabili per il monitoraggio della salute possono facilmente monitorare e fornire dati sulla frequenza cardiaca e sui livelli di attività dei pazienti ai servizi sanitari. Poiché la quantità di dati è enorme e proviene da molte fonti, le soluzioni basate sull'intelligenza artificiale vengono utilizzate per elaborare i dati e trovare anomalie per gli individui. Analogamente, negli ospedali, i dati generati dai dispositivi di monitoraggio della salute dei singoli pazienti possono individuare eventuali emergenze e allertare gli operatori sanitari. Alcuni Paesi, come la Norvegia e la Danimarca, stanno già utilizzando l'analisi del sistema sanitario per evidenziare

gli errori di trattamento e le inefficienze del flusso di lavoro. In questo modo, l'IA contribuisce a ridurre l'onere del sistema sanitario, evitando diagnosi errate e ricoveri non necessari e facendo risparmiare denaro e tempo ai pazienti, evitando appuntamenti inutili. L'addestramento basato sulla conoscenza dei modelli DL richiede insiemi di dati di input ottenuti dai dati degli studi clinici. Ad esempio, i dati di risonanza magnetica funzionale (fMRI) raccolti da pazienti affetti da Alzheimer e le scansioni di tomografia computerizzata (TC) del cancro ai polmoni sono utilizzati come file di input per la diagnosi assistita dall'intelligenza artificiale rispettivamente del morbo di Alzheimer e del cancro ai polmoni. La classificazione a livello di soggetto di 138 soggetti per gli stadi associati di AD ha portato a un'accuratezza del 100% per i soggetti cognitivamente normali (CN), del 96,85% per i disturbi soggettivi della memoria (SMC), del 97,38% per la compromissione cognitiva lieve precoce [*Early Mild Cognitive Impairment*] (EMCI), del 97,43% per la compromissione cognitiva lieve tarda [*Late Mild Cognitive Impairment*] (LMCI), del 97,40% per la compromissione cognitiva lieve [*Mild Cognitive Impairment*] (MCI) e del 98,01% per l'AD [35]. I dati di input raccolti nel corso degli anni aiutano anche a identificare i modelli nei dati. Un modello DL sviluppato da Etemadi et al. sulla base di 40.000 scansioni TC disponibili in precedenza ha superato i radiologi esperti identificando il cancro al polmone più precoce con un'accuratezza del 94%. [31]. Questi approcci contribuiranno all'individuazione dei tumori polmonari in fase precoce, cosa estremamente importante per l'assistenza sanitaria, poiché circa il 70% dei tumori polmonari viene individuato in fasi successive, con un conseguente basso tasso di sopravvivenza [13].

Analogamente, sono stati sviluppati modelli DL per l'analisi del cancro al seno [51,52] e del cancro al pancreas [32,53]. Il modello DL sviluppato da Alexander et al. ha raggiunto una sensibilità del 50,9% rispetto al 22,4% del metodo comunemente utilizzato basato sulla densità mammaria. Muhammed et al. e Qureshi et al. hanno sviluppato modelli DL basati su dati di scansione TC, ottenendo un'accuratezza rispettivamente

dell'80% [32] e dell'86% [53]. In altri casi l'IA prevede l'insufficienza cardiaca usando i dati dell'elettrocardiogramma (ECG) [54,55]. Usando solo i dati ECG come input per il modello DL, Akbilgic et al. hanno ottenuto un'area sotto la *Receiver Operating Charasteric Curve* (AUC) di 0,756 (0,717-0,795), che ha mostrato un'AUC ulteriormente migliorata di 0,818 (0,778-0,859) quando si è utilizzato l'output del modello ECG-AI, l'età, il sesso, la razza, l'indice di massa corporea, lo stato di fumatore, la malattia coronarica prevalente, il diabete mellito, la pressione sanguigna sistolica e la frequenza cardiaca come predittori [54]. Un altro modello DL sviluppato da Bagci et al. ha raggiunto un'accuratezza del 95% nella ricerca di macchie di cancro nelle scansioni TC, rispetto al 65% del tasso medio di accuratezza dei radiologi [30]. Il modello DL sviluppato ha aiutato a rilevare il cancro ai polmoni, dove i dati della TAC non avevano rilevato alcuna anomalia. Questo approccio contribuirà all'individuazione dei tumori polmonari in fase precoce [13], cosa estremamente importante per l'assistenza sanitaria, poiché circa il 70% dei tumori polmonari viene individuato in fasi successive, con conseguente basso tasso di sopravvivenza. Anche i modelli DL per l'analisi del cancro al pancreas sono stati sviluppati analizzando le immagini della TAC e altri dati clinici correlati. Esistono esempi simili di previsione dell'insufficienza cardiaca da parte dell'IA utilizzando i dati dell'elettrocardiogramma (ECG) [13]. Anche l'adozione dell'IA negli interventi di chirurgia robotica, in particolare nella chirurgia spinale, è oggetto d'interesse per le industrie sanitarie. I robot basati sull'IA possono analizzare i dati di precedenti interventi chirurgici per sviluppare nuovi metodi. Questi robot possono eseguire interventi chirurgici più accurati, riducendo i movimenti accidentali [25]. Oltre alla chirurgia spinale, l'IA trova applicazione anche nella chirurgia poco invasiva, negli interventi assistiti da robot e nell'assistenza post-operatoria, ad esempio nel calcolo dei tempi di recupero [56].

La commercializzazione dei sistemi basati sull'IA ha creato una piattaforma di trasformazione tecnologica chiamata "PathAI". Queste tecnologie basate sull'IA hanno dimostrato di

migliorare i risultati sanitari dei pazienti attraverso una diagnosi efficiente della patologia. Analogamente, *"PAGER"* è un'applicazione per la gestione della sanità che aiuta a trattare i pazienti fornendo raccomandazioni appropriate. L'IA è utile anche al settore dello sviluppo dei farmaci, in quanto individua possibili nuovi farmaci utilizzando la modellazione molecolare e i dati medici di addestramento. L'IA ha anche contribuito allo sviluppo di tecnologie per le interfacce uomo-macchina. Le tecnologie che prevedono delle interfacce uomo-macchina richiedono un sensore che generi dati di alta qualità e un algoritmo di IA con una potente capacità di analisi dei dati. Queste tecnologie si sono rivelate utili in campo medico. Abbiamo ad esempio degli studi su arti artificiali e sensori indossabili per raccogliere dati in tempo reale sul paziente [19,57,58]. Questa rassegna si concentra sull'esplorazione dei progressi nello sviluppo di piattaforme IoMT e AI assistite per il monitoraggio cardiaco, la diagnosi del cancro, gli interventi chirurgici, il monitoraggio dei diabetici e altre patologie correlate, come illustrato nella Figura 4.

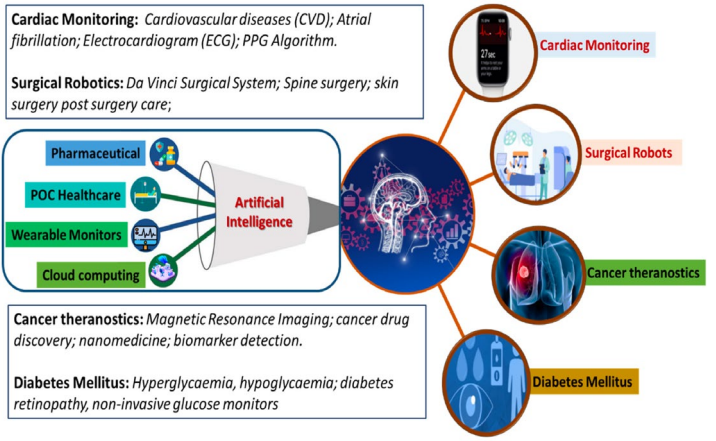


Figura 4. Rappresentazione schematica del ruolo degli approcci basati sull'IA in vari temi della ricerca sanitaria, tra cui il monitoraggio cardiaco, la chirurgia, la diagnostica del cancro e la gestione del diabete mellito.

[...]

4. MONITORAGGIO CARDIACO SUPPORTATO DA AI

Le malattie cardiovascolari (CVD), la principale causa di mortalità a livello mondiale, hanno causato quasi 18 milioni di vittime a livello globale solo nel 2019. Le CVD stanno diventando una grande preoccupazione, soprattutto nei Paesi a basso e medio reddito, a causa della loro allarmante morbilità. Oltre ai fattori genetici, anche lo stile di vita e l'alterazione data dallo stress diventano fattori di rischio significativi per le CVD umane. In base alle condizioni cliniche che influenzano la funzione del cuore, le CVD possono essere classificate in vari tipi, come insufficienza cardiaca, arresto cardiaco, infarti del miocardio, aritmia cardiaca, pericardite e cardiomiopatia. Per monitorare le irregolarità della funzione cardiaca sono disponibili diversi strumenti, tra cui la tomografia computerizzata (TC) cardiaca, l'elettrocardiogramma (ECG), il dispositivo di monitoraggio Holter, il test da sforzo e il profilo dei biomarcatori ematici. Tra questi, l'ECG, il test da sforzo e il profilo dei biomarcatori forniscono uno strumento rapido per una valutazione economica della funzione cardiaca. Lo stress è anche una delle principali cause di infarto del miocardio. Lo stress fisiologico è anche uno dei valori critici, che può essere monitorato in modo personalizzato a seconda dello stile di vita, delle situazioni e delle circostanze utilizzando un biosensore. I dispositivi di biosensing elettrochimici consentono di misurare lo stress stimando il cortisolo [72,96,97], un biomarcatore fisiologico dello stress [96].

Questa sezione riassume i recenti sviluppi degli strumenti di diagnosi rapida per individuare con precisione varie condizioni patologiche associate alle CVD. La fibrillazione atriale (FA) è una condizione di battito cardiaco irregolare e la popolazione affetta da FA è incline a sviluppare coaguli di sangue, che potrebbero portare a malfunzionamenti cardiovascolari e persino causare ictus. In questo caso si pone una sfida diagnostica perché la FA è

asintomatica e rimane inosservata fino a quando non si verifica il primo evento tromboembolico. Un evento tromboembolico è legato alla formazione di un coagulo di sangue nel vaso sanguigno, che il flusso sanguigno può trasportare fino a bloccare un altro vaso sanguigno. Sono stati condotti alcuni studi clinici per il monitoraggio continuo a lungo termine della fibrillazione atriale, come l'Embrace trial (uno screening di 30 giorni in pazienti con ictus criptogenetico) e il crystal AF trial (uno studio di 36 giorni di monitoraggio cardiaco continuo per valutare la fibrillazione atriale dopo un ictus criptogenetico). Tuttavia, come accennato in precedenza, questi studi presentano dei colli di bottiglia per tutti i pazienti con sospetta FA, a causa della scomodità di utilizzo, della mancanza di rimborsi e di altre ragioni tecniche. Per il monitoraggio a lungo termine della fibrillazione atriale, il metodo deve essere affidabile, economicamente vantaggioso, conveniente e dotato di strumenti di facile applicazione per il rilevamento esteso della fibrillazione atriale non invasiva.

L'elettrocardiogramma (ECG) è uno dei metodi contemporanei per il monitoraggio in tempo reale della funzione cardiovascolare. Per prevedere varie malattie cardiache, l'ECG è utile in quanto può fornire i dettagli morfologici e funzionali del cuore. L'irregolarità del ritmo cardiaco viene convenzionalmente monitorata utilizzando un registratore ECG a 12 derivazioni. Tuttavia, interrogare manualmente la registrazione ECG richiede molto tempo e l'analisi di un grande volume di dati può causare errori. Inoltre, lo strumento di misurazione ECG convenzionale causa disagi a chi lo indossa ed è soggetto a disturbi durante la misurazione ECG. Negli ultimi anni, i registratori ECG a canale singolo si sono evoluti come dispositivi sensibili per rilevare con precisione le variazioni del battito cardiaco. Tuttavia, i dati prodotti dal registratore ECG monocanale sono enormi e richiedono un programma automatizzato per elaborare il grande volume di dati e valutare la specificità della misurazione. I ricercatori hanno utilizzato algoritmi basati su ML per diagnosticare i battiti cardiaci aritmici e prevedere con precisione le anomalie. Dai dati ECG, le caratteristiche estratte possono essere utilizzate per rile-

vare condizioni cardiache come infarti del miocardio, tachicardia sinusale e apnea notturna [98]. Grazie ai progressi del cloud computing e alla capacità di elaborare un'ampia serie di dati, l'AI/ML si è dimostrata promettente nel monitoraggio dell'elettrofisiologia cardiaca e dell'imaging cardiaco. I sistemi basati sull'intelligenza artificiale (DL/ML) sono stati esplorati per varie applicazioni, tra cui l'analisi dei segnali ECG per la classificazione del rumore, l'identificazione delle aritmie, la previsione della fibrillazione atriale e l'analisi dell'intero genoma. Una panoramica del ruolo dell'AI/ML nelle misurazioni elettrofisiologiche è riportata nella Figura 8. I dati acquisiti dai dispositivi IoMT, come gli orologi intelligenti, le tecnologie dei telefoni cellulari e le immagini mediche, vengono addestrati ed elaborati dall'AI/ML per migliorare la diagnosi delle malattie, prevedere gli esiti e descrivere nuove patologie.

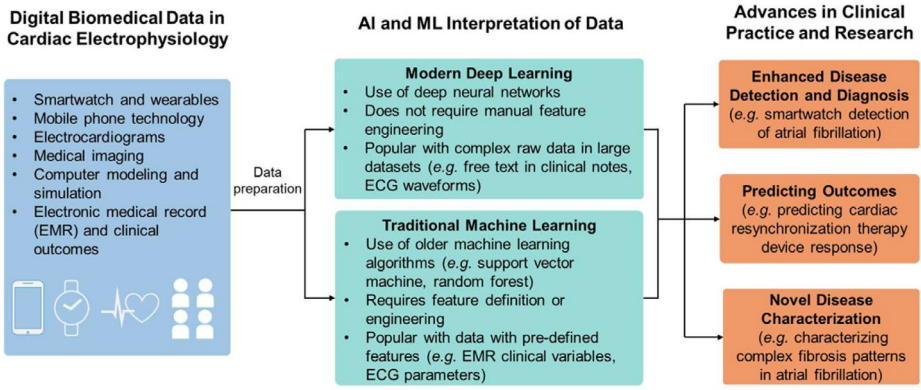


Figura 8. Ruolo dell'AI/ML in cardiologia. I dati biomedici raccolti attraverso la misurazione dell'elettrofisiologia cardiaca vengono interpretati attraverso algoritmi di ML tradizionali o moderni per migliorare i risultati.

Kachuee et al. hanno proposto un nuovo algoritmo di deep learning per l'analisi dei dati ECG, in grado di rappresentare il segnale in una forma adatta per la valutazione di diversi compiti, come il riconoscimento del segnale ECG e l'identificazione delle

irregolarità del battito cardiaco [99]. L'algoritmo di rete neurale profonda convoluzionale viene addestrato con Physionet MIT-BIH aritmia e il database diagnostico PTB. L'AI preelabora i segnali ECG prima di utilizzarli come input per la classificazione del battito cardiaco. È stata utilizzata una rete di convoluzione profonda per classificare il tipo di battito cardiaco dell'ECG e per il compito di previsione dell'addestramento. Poi è stata utilizzata la libreria tensor flow per l'addestramento e la valutazione del modello. La rete neurale a convoluzione profonda addestrata è stata utilizzata per valutare 4079 battiti cardiaci, per verificare se ci fosse aritmia. Tuttavia, il gruppo ha riferito che non è stato proposto un predittore esatto per i set di dati MIT-BIH, ma che il metodo pianificato ha mostrato un'ottima accuratezza rispetto ai metodi più avanzati. Gli smartphone basati sul fotopletismografo (PPG) sono utilizzati per lo screening della FA. Gli smartwatch con sensori PPG sono i metodi di nuova generazione adottati per rilevare la FA. Dörr et al. hanno proposto lo studio WATCH AF, confrontando l'accuratezza diagnostica nel rilevare la fibrillazione atriale da parte di un algoritmo PPG basato su smartwatch utilizzando i segnali PPG con la diagnosi dei cardiologi tramite ECG [100]. Il sensore PPG integrato nello smartwatch (Gear Fit 2) ha registrato i dati PPG raccolti dallo smartphone Samsung SE mini e li ha trasferiti al server. Un iECG SINGLE-LEAD (Alivecor Kardic Systems) è stato collegato a un iPhone 4s e i dati ECG sono stati raccolti e salvati in formato PDF. La seconda registrazione PPG è stata effettuata su un altro polso tramite una smart band (wavelet wristband) collegata a un Apple iPad mini e i dati sono stati trasferiti al server. I segnali PPG sono stati estratti come segmenti di 1, 3 e 5 minuti e proposti da un algoritmo PPG automatico.

In base al rapporto rumore-segnale, per l'analisi finale sono stati presi i dati di un segmento di 1 minuto. Confrontando la diagnosi basata sull'algoritmo del PPG con l'interpretazione dell'iECG da parte dei cardiologi, è emerso che la nuova diagnosi basata sull'algoritmo del PPG era efficace in termini di accuratezza complessiva del 92%, sensibilità del 98%, selettività

del 93,7% e migliori PPV (valore previsto positivo) e NPV (valore previsto negativo). Il limite principale del modello è il confronto solo con i dati dell'iECG, ma non con l'ECG standard a 12 derivazioni e l'ECG Holter. I dispositivi di rilevamento della FA basati su PPG possono ancora essere intesi come strumenti di screening della FA, con la necessità di un ECG di conferma per la sospetta FA. Li et al. hanno studiato i genomi per identificare l'aneurisma dell'aorta addominale (AAA) utilizzando un framework ML da genomi personali e cartelle cliniche elettroniche (EHR). In genere si tratta dell'effetto dei genomi personali e degli stili di vita individualizzati. Per la prima volta, hanno introdotto il sequenziamento dell'intero genoma (WGS) ad alta copertura per i pazienti con AAA con l'aiuto di HEAL (stima gerarchica da apprendimento agnostico). HEAL è un sottoinsieme che identifica i modelli distinti nei genomi e poi utilizza tali modelli per identificare gli esiti. Questo sottoinsieme identifica anche gli esiti a livello di mutazione. Si basa sulla struttura della stima gerarchica e dell'apprendimento agnostico [101].

5. IL RUOLO DELL'INTELLIGENZA ARTIFICIALE IN CHIRURGIA

L'importanza dell'IoMT sta rapidamente aumentando a causa della crescita dell'IA e dei suoi diversi sottoinsiemi, come il ML, la computer vision (CV), il deep learning (DL) e la processazione del linguaggio naturale (NLP). Tutti questi aspetti dell'IA forniscono una base per tutte le azioni autonome negli interventi chirurgici assistiti dall'IA [102]. La rapida crescita delle capacità dell'IA in campi come la chirurgia può essere attribuita alla combinazione di input e output dei sottocampi dell'IA. L'IA è applicabile a vari livelli della chirurgia. I dati collettivi sulle applicazioni dell'IA in diversi interventi di chirurgia spinale sono stati riassunti da Chang et al. [27]. Lo sviluppo dell'automazione negli interventi chirurgici ha portato a un maggiore utilizzo dell'IA. Negli interventi chirurgici tradizionali, l'uomo svolgeva tutti i ruoli. Nel corso del tempo gli interventi si sono completamente

affidati all'IA e si svolgono ora in autonomia, sia per ruoli parziali come la registrazione di immagini, sia per operazioni in cui non è richiesto il coinvolgimento diretto dell'uomo. Questo aspetto è stato discusso da Panesar et al. [103]. Il sistema chirurgico Da Vinci è uno dei più noti sistemi di chirurgia robotica assistita. Il sistema chirurgico consente ai medici di eseguire interventi da una cabina remota con tecnologie attrezzate per controllare i bracci del robot [104]. Si tratta di un metodo minimamente invasivo, che di solito gode della fiducia della maggior parte dei medici per la sua precisione e innovazione. Panesar e Ashkan et al. hanno discusso il ruolo di Internet o delle piattaforme mobili, che sono controllate dall'intelligenza artificiale e possono essere utilizzate per fornire competenze chirurgiche a distanza. Può essere utilizzato per guidare un robot chirurgico nell'esecuzione di interventi dove le risorse appropriate non sono disponibili o non è possibile accedervi, come ad esempio una navicella spaziale nello spazio o luoghi con disastri ambientali o guerre [105].

5.1 IL RUOLO DELL'INTELLIGENZA ARTIFICIALE NELLA CHIRURGIA SPINALE, CARDIACA E OCULISTICA

Di solito gli interventi chirurgici sono di molti tipi e alcuni possono essere programmati. Ciò non significa che l'intervento è facoltativo, ma che può essere programmato per comodità. Oppure c'è l'intervento d'emergenza, che può essere pericoloso per la vita a seconda delle condizioni mediche. Allo stesso modo, ci sono quattro obiettivi primari cruciali nella cura della chirurgia spinale, che includono (i) le pratiche preoperatorie, la selezione dei pazienti in base al loro livello di malattia e la previsione dei risultati dopo l'intervento; (ii) il miglioramento della qualità e della riproducibilità della ricerca spinale; (iii) la raccolta e il tracciamento dei dati prima dell'intervento; e (iv) le prestazioni chirurgiche intraoperatorie [104]. Questo massiccio aumento dell'intelligenza artificiale con l'apprendimento automatico apre ai chirurghi una nuova strada per analizzare i dati in modo più preciso [26]. L'intelligenza artificiale accompagnata dal ML viene

utilizzata per la diagnostica per immagini della colonna vertebrale, la previsione di interventi terapeutici, il recupero di informazioni, l'analisi biomeccanica e la caratterizzazione dei tessuti biologici (Figura 9).

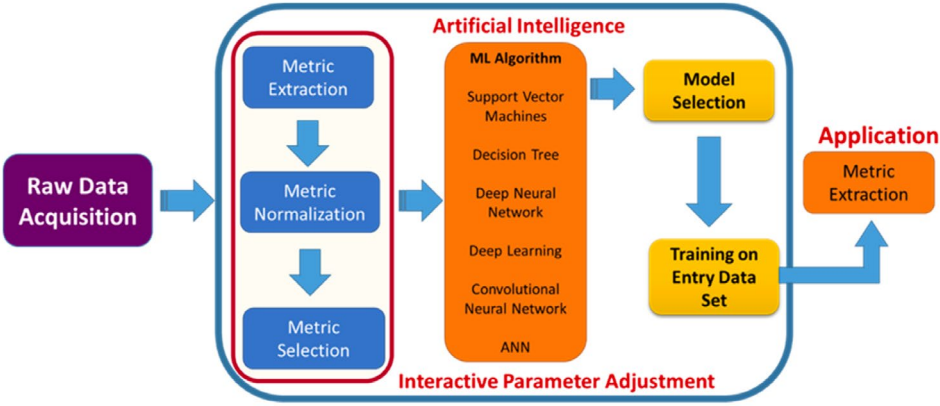


Figura 9. Ruolo dell'IA in chirurgia. Struttura che integra l'IA negli interventi di chirurgia spinale, che prevede l'acquisizione di dati grezzi per convertire gli input in forma digitalizzata e metodi di pre-elaborazione per l'apprendimento automatico, come l'estrazione di metriche per addestrare l'IA e la selezione di metriche per differenziare due gruppi. In base agli input viene selezionato l'algoritmo ottimale e quindi viene generato l'output.

A prescindere dal tipo di intervento, l'IA facilita le operazioni chirurgiche in molti modi. La spesa annuale per le cure spinali negli Stati Uniti è di circa 110 miliardi di dollari e si prevede che raggiungerà i 5,3 trilioni di dollari entro il 2025 (Rasouli et al., 2021). Inoltre, vi è un'enorme differenza nelle pratiche, nella fornitura di servizi, nell'assistenza e nei costi degli interventi chirurgici tra i diversi Paesi o addirittura tra i diversi ospedali dello stesso Paese. Ames et al. hanno suggerito che l'uso dell'IA in questi trattamenti chirurgici può aiutare a definire la qualità e le spese dei trattamenti e delle cure. Può anche migliorare i risultati e ridurre le spese dei pazienti e degli ospedali [106]. Sono stati proposti diversi modelli per consentire a pazienti e chirurghi di prevedere i rischi legati agli interventi chirurgici. Bekelis et al.

hanno usato un modello di regressione logistica per la valutazione del rischio [107]. Sheer et al. hanno usato un albero decisionale per la previsione delle complicanze negli interventi di deformità spinale negli adulti [108].

L'IA aiuta anche a raccogliere ed elaborare diverse informazioni, come i rischi connessi, le informazioni anatomiche, la genetica e altre storie della malattia e l'economia dei pazienti, e a fare migliori previsioni sugli interventi chirurgici [109]. Ad esempio in alcuni pazienti epilettici la malattia può essere prevista da un modello di DL, grazie al quale i pazienti potrebbero beneficiare meglio degli interventi chirurgici. L'intelligenza artificiale può fornire indicazioni ai chirurghi in sala operatoria in modo che l'intervento possa essere eseguito con rischi minimi. Alcuni studi precedenti hanno sviluppato degli algoritmi di apprendimento automatico in chirurgia cardiotoracica in grado di battere i punteggi di rischio operativi standard nel prevedere i decessi post-operatori nei pazienti cardiopatici [110]. Ostberg et al. hanno studiato come i metodi basati sul ML, come le reti neurali artificiali e le reti neurali convoluzionali, siano stati utilizzati in studi come la segmentazione dell'aorta ascendente e discendente e il rilevamento di patologie comuni ai raggi X del torace, il rilevamento di anomalie del movimento delle pareti sugli ecocardiogrammi e la segmentazione del ventricolo sinistro per misurare continuamente la frazione di eiezione negli interventi di chirurgia cardiaca/toracica [111]. Li et al. hanno discusso un approccio ML integrato con dati genomici e cartelle cliniche elettroniche, confermando una notevole capacità di studiare gli aneurismi dell'aorta addominale e i meccanismi genetici sottostanti [101]. Anche gli interventi di chirurgia cardiovascolare presentano il rischio di lesioni renali acute. I tassi di mortalità sono del 10,5% e del 30%, rispettivamente nel caso di interventi cardiaci e di lesioni renali acute, che aumentano con la gravità della lesione renale. Diversi metodi di apprendimento automatico aiutano a prevedere i rischi post-operatori di lesioni renali associate a interventi cardiaci. Modelli come la regressione logistica, l'albero decisionale semplice, la foresta casuale, la macchina vettoriale

di supporto, l'extreme gradient boosting e l'ensemble si sono dimostrati promettenti e hanno contribuito a ridurre al minimo le complicazioni postoperatorie [98].

Phillips et al. hanno effettuato una valutazione su un algoritmo di intelligenza artificiale per rilevare il melanoma in immagini di lesioni cutanee [112]. Rispetto ad altri tipi di cancro della pelle, il melanoma maligno è poco comune. Dovrebbe essere diagnosticato in fase iniziale e monitorato regolarmente. Una diagnosi di cancro allo stadio I ha un tasso di sopravvivenza relativa del 95% rispetto a una diagnosi tardiva allo stadio IV, con un tasso di sopravvivenza relativa compreso tra l'8% e il 25%. L'algoritmo utilizzato in questo caso è il DL. Zhu et al. hanno sviluppato una metodologia chimica analitica per ottenere un monitoraggio della pelle rapido, non invasivo e ad alto rendimento [113]. Per registrare il profilo di massa della superficie cutanea, viene combinata una procedura di campionamento per adesione con la spettroscopia di massa con desorbimento/ionizzazione laser assistito da matrice, a tempo di volo (MALDI-TOF [*Matrix-Assisted Laser Desorption/Ionization Time of Flight*]). I campioni leggeri di cellule sono raccolti con metodi di campionamento da strati epidermici della pelle, per aderenza. La spettroscopia di massa MALDI-TOF è messa in atto nei laboratori chimici e i risultati si ottengono in pochi minuti rilevando gli analiti in base al loro peso molecolare (Figura 10). È possibile inserire e analizzare più campioni alla volta. Lo spettro di massa può essere analizzato facilmente perché la maggior parte dei segnali è dovuta allo ione analita con carica singola. L'uso dell'intelligenza artificiale, come il ML, il data mining o l'analisi di reti complesse per l'interpretazione automatizzata dei dati, ci permette di elaborare rapidamente dati complessi ed estesi. Tuttavia, questo lavoro è ancora in fase sperimentale e non è ancora stato applicato alla pelle umana, ma è stato testato con successo nei topi e ha prodotto buoni risultati.

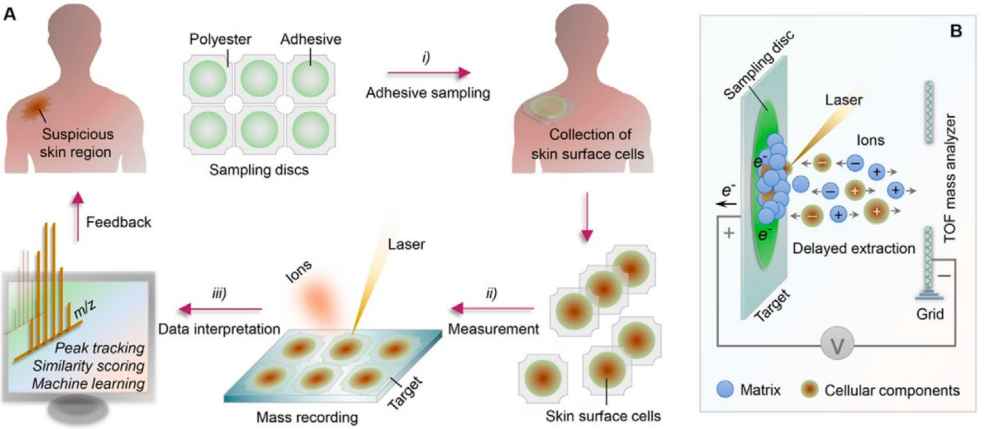


Figura 10. (A) Rappresentazione schematica della metodologia utilizzata per la diagnosi della pelle. (B) Rappresentazione della spettroscopia di massa con desorbimento/ionizzazione laser assistito da matrice, a tempo di volo (MALDI-TOF).

6. RUOLO DELL'IA NELLA GESTIONE DEL DIABETE MELLITO E DEL CANCRO

Il diabete rappresenta una preoccupazione crescente in ambito sanitario e rende essenziale il monitoraggio periodico dei livelli di glucosio nel sangue. La misurazione della glicemia è una delle principali per le persone affette da ipoglicemia e iperglicemia. A livello globale, circa 425 milioni di persone sono affette da diabete e circa il 12% della spesa totale mondiale è destinato alla gestione del diabete. Si stima che la spesa per il diabete aumenterà fino a ~490 miliardi di dollari entro il 2030. Il diabete cronico può portare alla retinopatia diabetica (DR), che causa cecità parziale o completa [114]. Secondo gli studi epidemiologici (cioè lo studio della frequenza e delle potenziali cause della malattia), una persona diabetica su tre soffre di DR, mentre la terza soffre di edema maculare diabetico, che è una piccola protuberanza che sporge dalle pareti dei vasi e perde sangue nella retina, causando gravi problemi di salute [14]. L'accuratezza e il fre-

quente monitoraggio del glucosio sono necessari per prevenire gli impedimenti clinici acuti e cronici causati dal diabete. Poiché la tecnica convenzionale di monitoraggio del glucosio richiede la puntura della pelle e il prelievo di sangue, è essenziale sviluppare una tecnologia per i pazienti a un costo accessibile senza pungere le dita più volte per controllare il livello di glucosio. I sistemi di monitoraggio del glucosio attualmente disponibili per la misurazione POC nei pazienti si basano su approcci elettrochimici. Sebbene siano presenti alcune carenze per quanto riguarda l'accuratezza e la precisione dei glucometri, il loro utilizzo per la gestione del diabete POC è in aumento ogni anno. Mentre i glucometri point-of-use forniscono un'istantanea delle tendenze del glucosio, un sistema di monitoraggio continuo del glucosio (CGM) fornisce informazioni in tempo reale sui livelli di glucosio sia al paziente che a chi lo assiste. La complessità delle dinamiche del sangue è una delle sfide più importanti per una previsione accurata e precoce dei livelli di glucosio. I metodi basati sull'AI/ML, sull'elaborazione del linguaggio naturale e sulle reti neurali artificiali sono molto importanti per il controllo del diabete, poiché aiutano a prevedere i modelli di diabete e a diagnosticarne il rischio, facilitando la gestione del diabete [115].

Approcci basati su AI e ML sono stati incorporati nei dispositivi di monitoraggio del glucosio per migliorare l'accuratezza clinica. Ad esempio, Khanam et al. hanno sviluppato un sistema combinato con l'intelligenza artificiale per rilevare con precisione i livelli di glucosio nell'uomo [38]. Per addestrare il sistema si usano cinque caratteristiche di input: età, gravidanza, indice di massa corporea, livelli di glucosio e insulina. Per valutare il set di dati sono stati utilizzati algoritmi ML come RF, NN, DT e K-nearest neighbor (KNN) con strati nascosti. Tutti i modelli hanno fornito una misura affidabile del glucosio con un'accuratezza superiore al 70%. Hamdi et al. hanno analizzato il livello di glucosio utilizzando un sistema ibrido con modelli compartimentali [116]. I livelli di glucosio sono stati monitorati utilizzando il metodo CGM con l'aiuto dell'algoritmo ANN. Il dispositivo consiste in un sensore sottocutaneo posizionato appena sotto la

pelle collegato al trasmettitore, che è ulteriormente collegato a un ricevitore wireless per visualizzare i livelli di glucosio. Il dispositivo misura i livelli di glucosio ogni 15 minuti. Per valutare il funzionamento del sistema nell'analisi clinica, sono stati raccolti i dati di 12 pazienti per le fasi di addestramento e convalida. L'algoritmo è composto da tre strati: uno strato di ingresso, uno strato nascosto e uno strato di uscita (Figura 11).

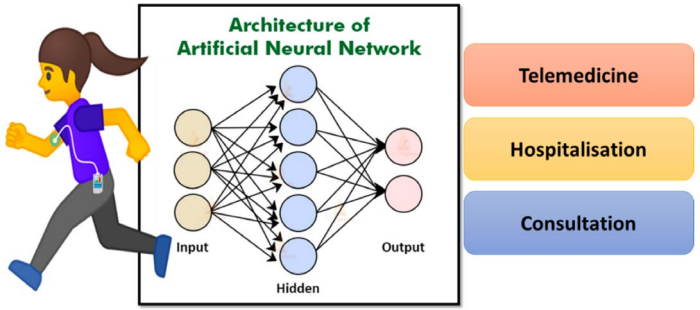


Figura 11. Rappresentazione della rete neurale artificiale che utilizza i dati del paziente per identificare il piano di trattamento corretto.

Rigla et al. hanno monitorato il diabete gestazionale con un gruppo di 247 pazienti utilizzando un sistema di supporto decisionale attraverso la telemedicina durante la pandemia di COVID-19 [117]. Il sistema di monitoraggio è costituito da uno smartphone per le informazioni sulle attività fisiche e da un glucometro con Bluetooth per trasferire i dati. I dati raccolti dal referto medico elettronico dell'ospedale (EMR), dal monitor della pressione sanguigna e dal glucometro vengono inviati all'algoritmo del sistema di supporto alle decisioni (DSS) come input per l'addestramento. L'algoritmo ML è programmato per suggerire al paziente un piano dietetico e indica al paziente se è necessaria una visita medica. La nutrizione di precisione personalizzata sta alla base dell'individualizzazione della nutrizione/dieta in base al tasso di metabolismo, al sesso, all'età e alla biochimica per il trattamento, la gestione e la prevenzione delle malattie. I dispositivi biosensoriali integrati consentono di sviluppare approcci perso-

nalizzati allo sviluppo dei nutrienti monitorando specifiche molecole biochimiche. Wang et al. hanno recentemente proposto il concetto di monitoraggio nutrizionale personalizzato integrando dispositivi di biosensoristica con sistemi basati su cloud [118]. La piattaforma di rilevamento multimodale proposta monitora l'assunzione e l'ingestione di cibo attraverso sensori di immagine e di movimento. Il sistema utilizza anche sensori indossabili per misurare i metaboliti e i nutrienti nei biofluidi umani, come sangue, sudore, saliva, urina e liquido interstiziale (ISF) (Figura 12).

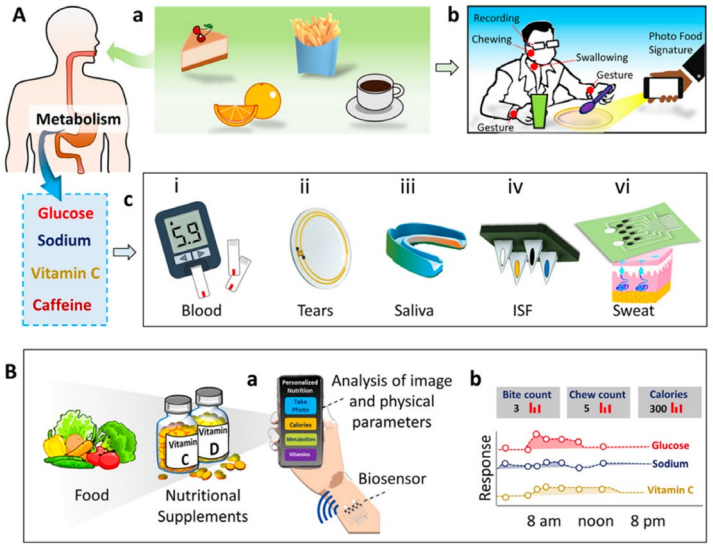


Figura 12. Concetto di sistema di misurazione della nutrizione personalizzato. (A) Monitoraggio dell'assunzione di cibo (a) e del comportamento di ingestione (b). Rilevamento indossabile dei metaboliti nel biofluido umano (c). (B) Rappresentazione schematica di un sistema completo di monitoraggio dei nutrienti per il monitoraggio simultaneo dei nutrienti presenti negli alimenti e dei metaboliti nell'uomo. (Riprodotto con il permesso dell'American Chemical Society) [118].

Marcus et al. hanno sviluppato un approccio SML personalizzato per la previsione del livello di glucosio. Le tecniche SML consentono di identificare modelli e relazioni negli insiemi di dati, che sono non lineari. Affrontare la non linearità è essenzia-

le per affrontare l'accuratezza clinica dei sensori di glucosio. Per addestrare il sistema sono stati raccolti i dati di 11 volontari. I dispositivi CGM utilizzano il liquido interstiziale (ISF) per valutare le concentrazioni di glucosio. Tuttavia, lo scarto temporale tra le variazioni di glucosio in entrambi i compartimenti è riportato tra i 5 e i 25 minuti. A causa di questo ritardo, le iniezioni automatiche di insulina (pancreas artificiale) non sono raccomandate per i CGM. In particolare, il CGM spesso non riesce a identificare in tempo gli eventi ipoglicemici. L'algoritmo sviluppato da questo team prevede gli aumenti dei livelli di zucchero con 30 minuti di anticipo, consentendo l'implementazione di sistemi di pancreas artificiale nei CGM [119].

Il rilevamento non invasivo del glucosio è un'area emergente per il monitoraggio continuo del glucosio e ha il potenziale per sostituire l'approccio convenzionale della puntura del dito. Si sono sperimentati dei sensori ottici basati su ML per il monitoraggio dei livelli di glucosio [24] che utilizzano sorgenti luminose di diverse lunghezze d'onda. Sono state utilizzate più di 21 sorgenti luminose con diverse lunghezze d'onda e sono stati implementati cinque diversi approcci ML per l'analisi e la previsione del glucosio. L'accuratezza della previsione del sistema è stata migliorata organizzando i set di dati in 21 classi. Il sistema ha dimostrato una buona capacità di discriminare tra livelli di glucosio più alti, più bassi e normali. Per la misurazione POC sono stati presentati dispositivi analitici su carta (μPAD) integrati con uno smartphone per il rilevamento colorimetrico dei livelli di glucosio nella saliva [120]. Sono state testate tre diverse combinazioni di agenti cromogeni per produrre colore quando il glucosio reagisce sulla superficie di lavoro dei μPAD. Le immagini sono state acquisite con quattro diversi smartphone in sette diverse condizioni di illuminazione. Sono stati esaminati diversi classificatori ML e i migliori classificatori per ogni condizione di rilevamento sono stati ottimizzati per migliorare l'accuratezza del rilevamento. I set di dati sono stati quindi elaborati utilizzando un sistema basato su cloud che controlla il classificatore in remoto.

Il cancro è la condizione clinica di crescita ingestibile e diffusione di cellule atipiche in tutto il corpo. I test di imaging utilizzati per individuare il cancro sono la tomografia computerizzata (TC), la risonanza magnetica (RM), la tomografia a emissione di positroni (PET) e gli ultrasuoni. Sono state sviluppate anche strategie di rilevamento basate su biomarcatori per individuare precocemente i marcatori proteici del cancro approvati dalla FDA [121,122]. Le tecnologie basate sull'IA hanno potenziali applicazioni nella ricerca sul cancro, tra cui la diagnosi precoce, lo screening in un'ampia popolazione, la classificazione e il grado di stadio, la caratterizzazione molecolare, la previsione degli esiti dei pazienti, le risposte ai trattamenti, il trattamento personalizzato, il flusso di lavoro automatizzato della radioterapia e la scoperta di nuovi farmaci antitumorali e le sperimentazioni cliniche. Gli algoritmi di intelligenza artificiale e di ML possono essere utilizzati anche per creare modelli di previsione per la valutazione delle metastasi linfonodali, della risposta ai trattamenti farmacologici e della prognosi. Utilizzando dati clinici, patologici e polimorfismi genetici in un modello ANN, i ricercatori hanno potuto prevedere lo stadio preoperatorio del cancro allo stomaco con un'accuratezza dell'82% [123]. Molti ricercatori hanno utilizzato algoritmi basati sull'intelligenza artificiale per sviluppare modelli computazionali per prevedere gli esiti del cancro. L'espressione genica si è dimostrata un metodo efficiente, ma presenta l'inconveniente di una dimensione limitata dei campioni. Il sistema ML accoppiato all'IA si è dimostrato efficiente nel rilevamento, nella diagnosi e nella classificazione dei sottotipi [79]. Gli algoritmi di ML sono stati riconosciuti come uno strumento alternativo preferito per il riconoscimento dei pattern nel cancro al seno [51]. Gli algoritmi di previsione decisionale, tra cui K-nearest neighbor (KNN), support vector machine (SVM) e decision tree (DT), aiutano a estrarre le caratteristiche cliniche dai set di dati molto ampi. Fan et al. hanno sviluppato un metodo immunoistochimico basato sulla 3,3'-diaminobenzidina (DAB) per individuare il multi-tumore [124]. Sulla base del modello proposto, il gruppo ha diagnosticato il tumore al seno iperespresso

da HER2 con un'elevata sensibilità (95%) e selettività (100%). L'AI ha aiutato il metodo immunoistochimico a rilevare il multi-tumore, superando la limitazione della sensibilità del metodo immunoistochimico manuale.

7. SFIDE E PROSPETTIVE FUTURE

Sono stati riportati i vantaggi dell'IA in diversi settori della sanità, come il monitoraggio dell'aritmia cardiaca, la gestione del diabete e gli interventi chirurgici assistiti. L'intelligenza artificiale aiuta a elaborare efficacemente dati sensoriali estesi e complessi per ulteriori analisi e migliorare le capacità decisionali. L'AI/ML aiuta anche a estrarre i dati analitici da set di dati a bassa risoluzione o rumorosi. Attraverso gli approcci SML, la tecnica AI/ML consente ai dispositivi IoMT di estrarre le informazioni nascoste in base alla relazione tra i parametri del campione e i segnali misurati. Le tecniche di AI migliorano anche l'intensità del segnale, la sensibilità, la specificità e il tempo di misurazione (Figura 13).

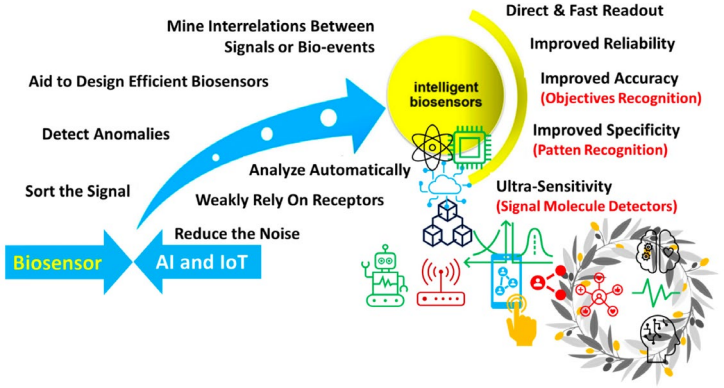


Figura 13. Ruolo dell'AI/ML nel miglioramento delle prestazioni dei sistemi biosensoriali. (Riprodotto con il permesso dell'American Chemical Society) [125].

Sebbene l'AI/ML abbia mostrato il potenziale per rivoluzionare la pratica sanitaria e i dispositivi medici integrati nell'IoMT,

è necessario affrontare numerose sfide tecnologiche per commercializzarla e calarla nelle cliniche e nella società. Poiché i sistemi AI/ML si basano molto sull'accuratezza dei dati per la programmazione e l'addestramento del sistema, l'attenzione deve essere rivolta alla raccolta di dati approfonditi sulla qualità dell'addestramento e dell'apprendimento dei pazienti. Un'altra sfida critica è l'eterogeneità dei dati raccolti. Le cartelle cliniche raccolte da diverse cliniche presentano vari tipi di distorsione e rumore, che causano discrepanze nell'addestramento dell'IA. Sofisticati algoritmi di ML possono aiutare a omogeneizzare i set di dati per migliorare l'accuratezza della diagnosi clinica. Con le tecniche supportate dall'IA, il futuro della chirurgia e del settore medico è destinato a fiorire.

La connettività delle tecnologie è fondamentale per collegare le persone con i dispositivi IoMT. La connettività può essere unidirezionale o bidirezionale. Il segnale del sensore IoMT deve essere elaborato prima di essere inviato al microcontrollore/processore, che richiede solo dati digitali. Il front-end analogico (AFE) elimina l'elettronica di massa necessaria per il condizionamento del segnale. La connettività tra l'AFE e il microcontrollore è solitamente stabilita da protocolli di comunicazione come I2C, SPI e UART. Tra l'AFE e il microcontrollore devono esserci dei protocolli di comunicazione compatibili. Wi-Fi e Bluetooth sono i principali metodi di connettività per far comunicare i dispositivi IoMT con l'hub centrale. La connettività Bluetooth è adatta per comunicazioni a corto raggio fino a 10 m con una velocità massima di 3 Mbps, che collega i sensori a vari gadget portatili come tablet, smartphone e PC. La trasmissione dei dati tramite Bluetooth è adatta soprattutto all'interno di sale operatorie, unità di terapia intensiva e altri luoghi con un numero maggiore di dispositivi. Per collegare i dispositivi IoMT al gateway si utilizza il Wi-Fi aziendale, che offre un livello di servizio superiore in termini di sicurezza e prestazioni. Quando il dispositivo di misurazione è in movimento, la connettività a Internet potrebbe essere interrotta, con conseguente perdita di dati critici e ritardo nella somministrazione dell'assistenza sanitaria al paziente. Per-

tanto, il modulo Wi-Fi deve supportare algoritmi di scansione ottimizzati per mantenere la rete per i dispositivi mobili in questi ambienti RF rumorosi.

I dispositivi medici convenzionali per la diagnostica richiedono componenti elettronici complessi e massicci per compensare gli errori elettrici e fisici. Ad esempio, i dispositivi ottici devono definire il percorso ottico e l'eliminazione delle interferenze della luce ambientale, il che richiede un design complicato dell'involucro. I dispositivi ottici progettati per le applicazioni portatili devono sempre rilevare la fluorescenza a bassa resa e allo stesso tempo eliminare il rumore del sistema. Sebbene i biosensori elettrochimici tradizionali richiedano stazioni di lavoro elettrochimiche, ingombranti e costose, i dispositivi POC o indossabili sono elettronica portatile. Possiamo sostituire l'elettronica richiesta su scala di massa con soluzioni a singolo circuito integrato, note anche come AFE. La maggior parte degli AFE è disponibile come pacchetto separato per i diversi tipi di sensori. Ad esempio, l'AD5940 di Analog Devices Inc. può essere utilizzato solo per biosensori elettrochimici. Questo AFE presenta limitazioni per quanto riguarda il multiplexing di diversi tipi di sensori, come i sensori ottici ed elettrochimici in un unico AFE. L'attuale generazione di dispositivi IoMT richiede AFE multifunzionali con più canali per interfacciarsi con una serie di sensori. Ad esempio, il potenziostato miniaturizzato [126] (M-P), sviluppato attraverso la personalizzazione dell'LMP91000, offre funzionalità di test POC e fornisce misure a basso consumo [127,128,129] ed elevata sensibilità. Tuttavia, l'interfacciamento multicanale dell'M-P e l'ulteriore funzionamento dello smartphone sono ancora impegnativi, ma presentano aspetti positivi.

I dispositivi di nuova generazione consentiranno anche un'elaborazione efficiente del segnale ricevuto, sia per la riduzione del rumore che per la gestione dei dati. La mobilità ha aperto nuove strade per i dispositivi sanitari IoMT. L'ambito della mobilità nell'IoMT è cresciuto anche con l'aumento dell'accesso agli smartphone. I fornitori di servizi sanitari stanno optando per dispositivi mobili basati su smartphone per un funziona-

mento semplice. Le tecnologie sanitarie mobili consentono la telecomunicazione, mettendo in contatto l'operatore sanitario e il paziente per una consultazione online a costi contenuti. I dispositivi IoMT dotati di applicazioni mobili possono controllare, accedere, monitorare e tracciare i dispositivi intelligenti. Tuttavia, i dispositivi mobili devono affrontare problemi di sicurezza e privacy. I fornitori di servizi sanitari stanno esplorando soluzioni innovative per affrontare le sfide della sicurezza e della trasmissione sicura dei dati.

Un altro problema critico associato all'IA è la gestione di una grande quantità di dati. Questi dati sono solitamente chiamati "big data". Questo problema si presenta quando vengono forniti input enormi e in tali condizioni il rischio di collasso del sistema è elevato. È quindi necessario un algoritmo di apprendimento automatico intelligente e scalabile per elaborare questi big data [130]. L'altra preoccupazione fondamentale è la privacy dei dati. Tutti i dati dei pazienti sono memorizzati nel cloud e c'è un'alta possibilità di leak e uso improprio dei dati [131]. Come qualsiasi altro trattamento o dispositivo medico, anche i dispositivi o i trattamenti basati sull'IA sono sottoposti a rigorosi processi di validazione prima di entrare nella pratica clinica. I processi di validazione e approvazione variano a seconda degli usi e delle forme previste. L'IA nell'assistenza sanitaria è progettata per svolgere un lavoro alla volta. L'ondata iniziale di sforzi viene compiuta per comprendere e risolvere i problemi di produzione. Uno dei problemi principali è la necessità di esperti in grado di risolvere i risultati complessi per gli esseri umani. Come discusso in precedenza, l'IA non è in grado di affrontare il multitasking. Per superare questo problema, non è ancora stato sviluppato un sistema in grado di lavorare in multitasking. L'influenza delle tecnologie AI sul settore sanitario sta migliorando e probabilmente dominerà nel prossimo decennio. Anche con i recenti progressi, la sostituzione del ruolo diagnostico del medico con i computer è ancora lontana.

Siamo nell'era dell'esplorazione della profondità dell'AI/ML nella diagnosi sanitaria, che prospererà nei sistemi IoMT e ci

aiuterà a sbloccare nuove strade nel settore sanitario. Con i progressi della nanotecnologia e della microelettronica, i dispositivi IoMT basati sull'intelligenza artificiale si muoveranno verso la funzionalità multilivello, l'alta sensibilità, la produzione a livello industriale, la miniaturizzazione, il bassissimo consumo energetico e l'economicità. Con la continua integrazione dei sistemi AI/ML con i dispositivi IoMT e la medicina integrata, le persone avranno progressivamente accesso a un'assistenza sanitaria di alta qualità.

8. CONCLUSIONI E OPINIONI

Questa rassegna riporta i recenti sviluppi e progressi fatti soprattutto nei dispositivi IoMT supportati dall'intelligenza artificiale per un biosensing che sia più efficiente nella gestione delle malattie. A sostegno degli aspetti dell'IA e dell'IoMT, abbiamo discusso l'importanza delle nanotecnologie nella piattaforma IoMT per lo sviluppo di dispositivi biomedici di ultima generazione come e-skin, e-nose ed e-textiles. I dispositivi IoMT integrati con l'intelligenza artificiale servono in aree mediche importanti, come il monitoraggio cardiaco, gli interventi chirurgici, il diabete e il monitoraggio del cancro. Grazie al cloud computing, l'intelligenza artificiale si è dimostrata promettente nel monitoraggio dell'elettrofisiologia cardiaca e della diagnostica per immagini. Sono degne di nota le innovazioni apportate dall'IA nel settore della chirurgia. Una di queste notevoli invenzioni è Davinci, un sistema chirurgico assistito da robot. L'IA ha generato un grande cambiamento nella gestione personalizzata del diabete e del cancro.

I risultati basati sull'intelligenza artificiale supportano la predizione anticipata e determinano il livello di rischio durante la diagnosi della malattia. Molti medici preferiscono gli algoritmi ML per la predizione, grazie all'accuratezza dei dati. Oltre agli algoritmi di ML, anche sottoinsiemi di IA come SVM e NN sono ampiamente utilizzati nel settore clinico. Ogni innovazione pre-

senta delle sfide, ad esempio l'IA deve affrontare problemi di eterogeneità, connettività e gestione dei dati. Le soluzioni a questi problemi sono state discusse in questa rassegna. A quanto pare, l'IA non può svolgere più lavori contemporaneamente e ancora non si è riusciti nella sostituzione completa del medico. Nonostante questa lacuna, l'IA ha dimostrato la sua eccellenza in campo medico grazie a una continua evoluzione. Il risultato di questa rassegna motiva i giovani ricercatori a promuovere e studiare approcci combinati che coinvolgano il rilevamento nano-abilitato, l'IA e l'IoMT per un biosensing efficiente, necessario per il controllo e la gestione delle malattie in modo personalizzato.

5 - UN SANO DIBATTITO: ESPLORARE LE OPINIONI DEI MEDICI SULL'ETICA DELL'INTELLIGENZA ARTIFICIALE

Tratto e tradotto da
Andreia Martinho, Maarten Kroesen, Coro Caspar (2021) A healthy debate: Exploring the views of medical doctors on the ethics of artificial intelligence. In (Ed.), Artificial Intelligence in Medicine- Volume 121, Novembre 2021, 102190.

https://doi.org/10.1016/j.artmed.2021.102190

Le parti omesse dal curatore rispetto all'originale sono indicate dal segno **[...]**

HIGHLIGHTS

- I medici hanno opinioni diverse sull'etica dell'IA sanitaria.
- La maggior parte dei settori dell'assistenza sanitaria può beneficiare dell'IA.
- I medici hanno una forte opinione sul ruolo delle grandi aziende nell'assistenza sanitaria.
- L'IA deve essere allineata ai principi bioetici fondamentali
- I medici devono partecipare al processo di progettazione dell'IA.

ABSTRACT

L'intelligenza artificiale (IA) si sta muovendo verso lo spazio sanitario. È generalmente riconosciuto che, pur essendo molto promettente l'implementazione delle tecnologie AI nell'assisten-

za sanitaria, essa solleva anche importanti questioni etiche. In questo studio abbiamo intervistato medici con sede nei Paesi Bassi, in Portogallo e negli Stati Uniti, provenienti da un mix eterogeneo di specializzazioni mediche, in merito all'etica che circonda l'IA in ambito sanitario. Dai dati sono emerse quattro prospettive principali che rappresentano diversi punti di vista sulla questione. La prima prospettiva (L'IA è uno strumento utile: lasciate che i medici facciano ciò per cui sono stati formati) sottolinea l'efficienza associata all'automazione, che consentirà ai medici di avere il tempo di concentrarsi sull'ampliamento delle loro conoscenze e competenze mediche. La seconda prospettiva (Norme e regolamenti fondamentali: Le aziende private pensano solo ai soldi) mostra una forte sfiducia nelle aziende tecnologiche private e sottolinea la necessità di una supervisione normativa. La terza prospettiva (L'etica è sufficiente: ci si può fidare delle aziende private) ripone una maggiore fiducia nelle aziende tecnologiche private e sostiene che l'etica è sufficiente a fondare queste società. Infine, la quarta prospettiva (Strumenti di IA spiegabili: l'apprendimento è necessario e inevitabile) sottolinea l'importanza della spiegabilità degli strumenti di IA per garantire che i medici siano coinvolti nel progresso tecnologico. Ciascuna prospettiva fornisce spunti preziosi e spesso contrastanti sulle questioni etiche che dovrebbero essere rese operative e tenute in considerazione nella progettazione e nello sviluppo dell'IA sanitaria.

1. INTRODUZIONE

L'Intelligenza Artificiale (IA) si sta muovendo verso lo spazio sanitario. Data l'abbondanza di dati generati dai sistemi sanitari come risultato degli sforzi di digitalizzazione compiuti nell'ultimo decennio, è emerso un nuovo approccio basato sui dati per implementare l'IA in ambito sanitario. A differenza dei precedenti approcci, in parte falliti, basati su regole per implementare l'IA nell'assistenza sanitaria [1,2], questo nuovo approccio si basa

fortemente su algoritmi che rilevano modelli nei dati provenienti dalla pratica clinica (ad esempio, imaging medico e cartelle cliniche elettroniche), dalle sperimentazioni cliniche, dagli studi di genomica e dalle operazioni di gestione dei benefici assicurativi, farmaceutici e delle farmacie [3]. Si prevede che questi metodi e algoritmi di IA all'avanguardia, basati sui dati, saranno in grado di utilizzare tali dati per affrontare i complessi problemi dei sistemi sanitari [4,3].

L'implementazione dell'IA nell'assistenza sanitaria è molto promettente per ampliare le conoscenze mediche e fornire soluzioni sanitarie ottimali ed economicamente vantaggiose [5,6]. In ambito clinico, i risultati attesi includono l'identificazione di individui ad alto rischio di malattia, il miglioramento della diagnosi e l'abbinamento di trattamenti personalizzati efficaci e il monitoraggio extraospedaliero della risposta alla terapia [4,7]. Nonostante i benefici previsti, l'IA sanitaria solleva anche importanti questioni etiche [8,9].

È noto che l'IA può potenzialmente minacciare valori come l'autonomia, la privacy e la sicurezza [10], che sono valori fondamentali in medicina [11,12]. Pertanto, affinché l'IA promuova la qualità delle cure e riduca al minimo gli effetti potenzialmente dirompenti [13], il suo impiego deve tenere conto dell'etica. Un passo importante verso l'impiego etico delle tecnologie di AI dirompenti è quello di conoscere il punto di vista degli operatori su tali tecnologie. Queste informazioni consentono una migliore operazionalizzazione delle questioni etiche associate all'IA in un particolare ambito, che alla fine dovrebbe portare a dibattiti più significativi e a politiche più solide.

L'attuale letteratura accademica fornisce informazioni interessanti e preziose sulle prospettive dei professionisti in merito all'impatto delle tecnologie AI nella professione medica [14,15,16,17]. La maggior parte di questi studi si riferisce in particolare a settori medici con una forte componente di elaborazione delle immagini, adeguata all'analisi automatizzata, come la radiologia [18,19,20,21,22,23,24], la patologia [25] e la dermatologia [26,27]. Tuttavia, le conoscenze sulle opinioni dei medi-

ci in merito alle questioni etiche associate all'implementazione dell'IA nell'assistenza sanitaria sono scarse.

L'obiettivo di questo studio è quello di comprendere i modelli di ragionamento e le opinioni morali sull'IA in ambito sanitario da parte di coloro che sono coinvolti nella pratica medica. Attraverso un sondaggio condotto tra medici olandesi, portoghesi e statunitensi sulle questioni etiche associate all'implementazione dell'IA nell'assistenza sanitaria, ci aspettiamo di arricchire la letteratura esistente sull'impatto delle tecnologie IA in medicina e di fornire conoscenze preziose per l'operatività dell'etica dell'IA sanitaria.

Per prima cosa forniamo un breve commento sull'etica dell'IA in ambito sanitario. Successivamente, illustriamo i metodi utilizzati in questa ricerca, delineando le fasi fondamentali della metodologia-Q e spiegando come abbiamo stabilito queste fasi in questo studio. Successivamente presentiamo i risultati dello studio descrivendo le quattro diverse prospettive emerse dai dati. Questi risultati vengono ulteriormente analizzati e discussi. Infine, tracciamo le conclusioni e presentiamo le direzioni per ulteriori ricerche.

2. L'ETICA DELL'IA PER LA SALUTE

Il lavoro empirico sull'IA nell'assistenza sanitaria che è stato riportato in letteratura si concentra principalmente su questioni direttamente correlate alla pratica e alla carriera medica, come il futuro dell'occupazione, la formazione sull'IA e la responsabilità.

È stato riportato che gli studenti di medicina e i medici comprendono la crescente importanza dell'IA nell'assistenza sanitaria e hanno atteggiamenti positivi verso l'uso clinico dell'IA [20,26,17], ma soprattutto come sistema di supporto alla diagnosi [18,19,26,27,25,24].

Nonostante l'atteggiamento positivo nei confronti dell'IA, è stato anche segnalato che gli studenti e i medici sono scarsamente formati su queste tecnologie [20,28,29,30]. Uno studio ha

indicato che, sebbene una piccola coorte di studenti di medicina del Regno Unito che avevano ricevuto un insegnamento sull'IA si sentissero più sicuri di lavorare con l'IA in futuro rispetto agli studenti che non avevano ricevuto un insegnamento, un numero significativo di studenti istruiti si sentiva ancora non adeguatamente preparato [20]. Per trarre il massimo vantaggio da queste tecnologie, gli studiosi sembrano concordare sulla necessità di ampliare e migliorare la formazione delle scuole di medicina sull'IA [18,20,21,26,25].

Per quanto riguarda l'impatto dell'IA sulle scelte di carriera e sulla reputazione, è stato riportato che l'IA ha un impatto sulle intenzioni di carriera degli studenti rispetto alla radiologia [20], ma i radiologi sceglierebbero comunque questa specialità se potessero scegliere [21]. Questi specialisti hanno tuttavia espresso il timore che l'IA possa diminuire la loro reputazione professionale [24].

Contrariamente alla percezione dell'opinione pubblica che l'IA sostituirà completamente o parzialmente i medici umani [31], gli studenti di medicina e i medici in generale non sono preoccupati della sostituzione del lavoro [18,26,17,32,24].

Un'altra questione importante legata alla pratica e alla carriera medica è la responsabilità. In uno studio in cui sono stati intervistati dei patologi, è stato riportato che, per quanto riguarda la responsabilità medico-legale per gli errori diagnostici commessi da una combinazione uomo/AI, le opinioni erano divise tra coloro che ritenevano che il fornitore della piattaforma e il patologo dovessero essere ritenuti ugualmente responsabili, e altri che ritenevano che la responsabilità rimanesse principalmente dell'uomo, con solo una minoranza che riferiva che il fornitore della piattaforma dovesse essere principalmente responsabile [25].

Chiaramente, l'etica che circonda l'implementazione dell'IA nell'assistenza sanitaria va oltre le questioni legate alla pratica e alla carriera medica. L'IA in campo sanitario solleva questioni etiche di livello superiore come l'autonomia, l'equità o la privacy [33,10] ma, ad eccezione dell'equità, questi temi hanno ricevuto meno attenzione nella letteratura scientifica. I problemi di equità legati ai pregiudizi razziali e di genere nelle applicazioni mediche

basate sull'IA hanno a che fare con il fatto che gli algoritmi di IA vengono addestrati su dati di pazienti bianchi prevalentemente di sesso maschile. Sia nella letteratura popolare che in quella scientifica sono state sollevate preoccupazioni sul fatto che questi algoritmi perpetuino e amplifichino i pregiudizi e le disuguaglianze esistenti nell'assistenza sanitaria [34,35,36,37,38]. È stato sottolineato che i dati medici devono essere valutati criticamente per evitare tali pregiudizi [34].

In questo studio empirico abbiamo intervistato i medici su una più ampia gamma di questioni etiche relative all'IA in ambito sanitario. Abbiamo affrontato i temi della privacy, dell'equità, della responsabilità, della trasparenza, della sicurezza, della supervisione umana, della spiegabilità, del futuro dell'occupazione, del finanziamento responsabile della ricerca, dell'educazione all'IA, dell'autonomia umana, della certificazione dei prodotti di IA e della progettazione etica. L'ampia gamma di temi etici dell'IA in campo sanitario esaminati in questo studio empirico ci permette di discernere i punti di vista e le opinioni morali dei medici sull'implementazione dell'IA in campo sanitario.

3. METODI

3.1 PANORAMICA

In questa ricerca abbiamo utilizzato la metodologia q, un approccio empirico sistematico per identificare le prospettive eventualmente conflittuali degli individui (stakeholder) su un particolare argomento [39,40,41,42,43]. La premessa fondamentale della metodologia-Q è che la soggettività è sempre autoreferente, cioè solo l'individuo può misurare la propria soggettività relazionale, cioè il significato di un'affermazione deriva dalla sua relazione con altre affermazioni, e si può dimostrare che ha una struttura e una forma [41]. Questo metodo è quindi considerato adeguato per il nostro scopo di discernere e studiare sistematicamente le opinioni soggettive dei medici sull'IA.

La metodologia Q richiede ai partecipanti di ordinare un insieme predefinito di elementi in base a una nozione soggettiva di accordo/disaccordo. In questo studio, i medici sono stati invitati a classificare una serie di affermazioni tratte dalla letteratura scientifica e popolare, che affrontano le principali questioni etiche dell'IA in una distribuzione a campana che va da -5 a +5, e a fornire ulteriori commenti sulle affermazioni che hanno classificato come più alte (+5) e più basse (-5). 1 Utilizzando tecniche statistiche, si sono formati gruppi coerenti che presentano prospettive particolari sull'etica dell'IA sanitaria. Interpretiamo queste prospettive e discutiamo in che modo si relazionano e si differenziano l'una dall'altra.

L'utilizzo della metodologia-Q presenta grandi vantaggi rispetto ad altri metodi di ricerca esplorativa, come interviste, focus group e sondaggi. A differenza delle interviste, gli studi q forniscono risultati numerici a sostegno delle prospettive soggettive su un particolare argomento, combinando così approcci quantitativi e qualitativi [44]. Inoltre, poiché i partecipanti ai Q-studies selezionano gli item individualmente, questi studi sono meno influenzati dagli effetti di dominanza, che si osservano in altri metodi di ricerca somministrati in gruppo, come i focus group [44]. Inoltre, a differenza dei sondaggi standard, in cui le opinioni dei partecipanti su ciascun argomento vengono estratte separatamente, i Q-studies richiedono ai partecipanti di considerare tali argomenti simultaneamente, scoprendo così connessioni latenti e consentendo opinioni più sfumate e sofisticate [44,45].

Per gli scopi del nostro studio, che ricordiamo è quello di rivelare i diversi punti di vista sull'etica dell'IA sanitaria, abbiamo anche ritenuto che la metodologia q fosse un metodo di ricerca più adatto rispetto al metodo Delphi [46]. Quest'ultimo è tipicamente utilizzato per la consultazione di esperti e in questo senso è simile al nostro studio q-metodologico, in cui intervistiamo i medici sull'IA sanitaria. Tuttavia, il metodo Delphi si concentra sul raggiungimento della convergenza (riducendo l'eterogeneità) tra gli esperti su determinati risultati incerti, mentre il metodo q

si concentra sulla rivelazione dell'eterogeneità tra gli stakeholder o gli esperti.

Questo studio ha seguito la tipica sequenza di quattro fasi degli studi q-metodologici che comprende (i) la *definizione del discorso comunicativo*; (ii) lo *sviluppo dell'insieme di affermazioni (Q-set)*; (iii) la *selezione dei partecipanti (P-set)*; (iv) l'*analisi e l'interpretazione*. Di seguito forniamo ulteriori dettagli su ciascuna di queste fasi in questo particolare studio.

3.2 IL DISCORSO DELLA COMUNICAZIONE

Il discorso della comunicazione è un corpus di opinioni relative a un particolare argomento [47]. Tali opinioni possono essere raccolte attraverso fonti dirette, come interviste e tecniche di gruppo nominale, o indirette, come articoli, forum di discussione e blog. In questo studio abbiamo utilizzato fonti indirette piuttosto varie, tra cui pubblicazioni scientifiche, pubblicazioni di divulgazione scientifica, associazioni professionali, società di consulenza e anche blog.

Abbiamo esaminato la letteratura scientifica e grigia sull'IA per la salute utilizzando le combinazioni di parole chiave "Intelligenza artificiale", "Machine Learning" e "Augmented Intelligence" insieme al connettore "AND" e alle parole chiave "Sanità", "Medicina", "medici" e "medici" in Google, Google Scholar e Web of Science.

Inizialmente abbiamo selezionato 353 affermazioni per la nostra comunicazione e successivamente le abbiamo assegnate a quindici cluster, utilizzando come strumento di orientamento un elenco di questioni etiche compilato da 22 importanti linee guida sull'etica dell'IA [10]. Ogni gruppo di affermazioni è stato associato a una particolare questione etica dell'IA tratta da tale elenco, ovvero Privacy; Equità; Responsabilità; Trasparenza; Sicurezza e Cybersicurezza; Supervisione umana; Spiegabilità; Futuro dell'occupazione; Finanziamento responsabile della ricerca; Educazione all'IA; Autonomia umana; Certificazione dei prodotti di IA; Progettazione etica; Deliberazioni specifiche per

la salute; è stato inoltre aggiunto un ulteriore gruppo riguardante l'IA nella pandemia Covid-19.

L'organizzazione del complesso di comunicazioni in gruppi che corrispondono a questioni etiche generali dell'IA ha facilitato la definizione dell'insieme q, poiché le affermazioni di questo insieme dovrebbero riflettere l'intero spazio delle questioni etiche identificate nel complesso. Va tuttavia sottolineato che l'elenco di questioni etiche utilizzato in questa ricerca come strumento di riferimento riflette un particolare approccio all'etica basato sulla deontologia. Altri approcci e principi etici potenzialmente rilevanti legati, ad esempio, al consenso informato e all'accettazione del rischio non sono quindi inclusi in tale elenco [48]. La ricerca futura potrebbe identificare ed esplorare ulteriormente ulteriori questioni e valori etici relativi all'IA sanitaria.

3.3 INSIEME DI AFFERMAZIONI (Q-SET)

L'insieme q è un sottoinsieme completo ma gestibile del discorso della comunicazione. Abbiamo analizzato ogni affermazione nei cluster definiti all'interno del discorso al fine di selezionare gli elementi rilevanti per ottenere un insieme strutturato, completo ed equilibrato di affermazioni. Questa selezione è stata guidata da tre considerazioni principali: (i) tenere conto di un'ampia gamma di posizioni presentate nella letteratura scientifica e divulgativa sull'AI Health; (ii) favorire la chiarezza; (iii) evitare la ridondanza. Utilizzando questo metodo di ottenimento del discorso, abbiamo cercato di ottenere il massimo dell'obiettività e della neutralità.

Il q-set finale comprende 40 affermazioni. Sono state apportate piccole modifiche a queste affermazioni per garantire la neutralità e per rispettare il numero di caratteri consentito da FlashQ [49], lo strumento software utilizzato in questo studio per la somministrazione del sondaggio. La dimensione dell'insieme è in linea con le pratiche correnti della metodologia q [41].

Il panorama delle affermazioni del q-set rispetto ai cluster etici predefiniti è composto da Privacy (affermazioni 1-4); Equi-

tà (5-8); Responsabilità (9-10,40); Trasparenza (11); Sicurezza e Cybersicurezza (12-13,39); Supervisione umana (18); Spiegabilità (15-17); Futuro dell'occupazione (19-20, 22); Finanziamento responsabile della ricerca (23-24); Educazione all'IA (25,34); Autonomia umana (18); Certificazione dei prodotti di IA (29-30); Progettazione etica (31-33); Deliberazioni specifiche per la salute (14,21,26-27,36-38); IA nella pandemia Covid-19 (28,35). L'insieme definitivo è elencato di seguito.

1. La privacy non dovrebbe essere la priorità assoluta nell'assistenza sanitaria basata sull'intelligenza artificiale.
2. La riservatezza non deve limitare l'implementazione dell'IA nell'assistenza sanitaria.
3. Senza regole chiare sull'utilizzo, l'archiviazione e l'anonimizzazione dei dati, l'IA non dovrebbe mai essere utilizzata in ambito sanitario.
4. La riservatezza, così come viene definita oggi, ha poca utilità in un futuro in cui l'assistenza sanitaria si affida in larga misura all'intelligenza artificiale.
5. È più probabile che l'IA risolva piuttosto che amplificare le disuguaglianze nell'assistenza sanitaria.
6. Migliorare l'equità e l'inclusione dovrebbe essere la priorità assoluta nello sviluppo e nell'impiego dell'IA nell'assistenza sanitaria.
7. L'intelligenza artificiale aumenterà la discriminazione basata sulla previsione di problemi medici futuri.
8. Dovremmo essere prudenti nel promuovere l'IA nell'assistenza sanitaria a causa delle questioni etiche irrisolte.
9. Gli sviluppatori di IA devono essere vincolati dall'etica medica.
10. In nome del progresso tecnologico, le aziende di IA non dovrebbero essere responsabili degli errori medici.
11. Gli strumenti medici di IA dovrebbero essere utilizzati solo se i medici comprendono come vengono prese le decisioni di IA.

12. Esiste un elevato rischio di comportamento monopolistico da parte di aziende private di IA nel settore dell'assistenza sanitaria.
13. Non è auspicabile che le grandi aziende entrino nel settore sanitario perché conoscono poco la medicina.
14. Il rapporto paziente-medico cambierà radicalmente una volta che l'IA sarà pienamente diffusa nei sistemi sanitari.
15. Gli operatori sanitari non hanno bisogno di sapere come funzionano gli strumenti medici dell'IA, ma piuttosto se sono affidabili.
16. I professionisti della salute si sono sempre fidati delle scatole nere (ad esempio la risonanza magnetica) e non sarà diverso con l'IA.
17. Non è possibile ottenere un consenso informato adeguato se il medico non è in grado di spiegare al paziente come funziona il dispositivo medico di IA.
18. L'IA diminuirà l'autonomia e l'autorità dei medici.
19. L'IA non sostituirà i medici, ma i medici che utilizzano l'IA sostituiranno i medici che non la utilizzano.
20. Se gli strumenti di IA funzionano bene, gli ospedali dovrebbero risparmiare assumendo personale meno qualificato.
21. L'intelligenza artificiale peggiorerà i problemi dell'assistenza sanitaria, come l'eccesso di test, di diagnosi e di trattamenti.
22. L'automazione può funzionare bene nelle fabbriche, ma non negli ospedali.
23. I prodotti medici basati sull'intelligenza artificiale non saranno all'altezza del clamore suscitato.
24. Tutti i fondi stanziati per l'IA sono utili se questa riesce a conquistare le sponde burocratiche, come la presa di appunti, la codifica e la ricerca di modelli.
25. I medici non sono interessati a conoscere l'IA e l'informatica.
26. In campo medico è problematico che le macchine manchino di conoscenze contestuali e della capacità di leggere gli indizi sociali.

27. Non sarebbe etico non utilizzare gli strumenti di intelligenza artificiale se questi forniscono decisioni migliori rispetto ai medici.
28. L'intelligenza artificiale ha già svolto un ruolo fondamentale nella pandemia COVID-19.
29. Il mantra dell'industria tecnologica "fallire velocemente e rimediare più tardi" sta mettendo a rischio i pazienti e le autorità di regolamentazione non stanno facendo abbastanza per garantire la sicurezza dei consumatori.
30. I prodotti sanitari di IA devono essere testati in studi clinici randomizzati, che sono la fonte più solida di prove mediche.
31. Poiché i sistemi di intelligenza artificiale sono progettati principalmente per aumentare i profitti, in futuro i sistemi sanitari disporranno di maggiori risorse e forniranno cure migliori.
32. La tecnologia AI in ambito sanitario deve essere allineata ai principi bioetici.
33. I medici devono partecipare al processo di progettazione dell'IA per l'assistenza sanitaria.
34. I medici non hanno il tempo di imparare a utilizzare i complessi dispositivi medici basati sull'intelligenza artificiale.
35. L'intelligenza artificiale migliora il processo decisionale medico in situazioni di razionamento delle cure.
36. L'intelligenza artificiale consentirà ai fornitori, ai medici e al personale di concentrarsi sulle attività e sulle competenze più importanti.
37. La maggior parte dei settori dell'assistenza sanitaria può beneficiare dell'IA.
38. Non è molto difficile rendere operativa la pratica clinica per una macchina.
39. La medicina non dovrebbe mai affidarsi all'intelligenza artificiale perché questi sistemi informatici sono vulnerabili alle minacce alla sicurezza informatica.
40. Se un medico commette un errore in seguito ai consigli di uno strumento di intelligenza artificiale, dovrebbe essere considerato responsabile.

3.4 PARTECIPANTI (Q-SET)

Per il reclutamento dei partecipanti a questo studio sono stati utilizzati tre approcci diversi, che hanno comportato il raggiungimento di: (i) dipartimenti ospedalieri (per telefono e successivamente via e-mail); (ii) medici che sono conoscenti personali; e (iii) medici che non sono conoscenti personali (attraverso indirizzi e-mail resi disponibili in pubblicazioni trovate in Google Scholar relative a vari campi della medicina). Dato che l'approccio (iii) si è rivelato molto più efficace, alla fine le altre strategie sono state abbandonate e ci siamo concentrati principalmente sul raggiungimento dei medici attraverso le pubblicazioni di cui erano autori di recente.

Piuttosto che concentrarci su un particolare settore medico, in cui gli operatori possono condividere pensieri simili sull'IA, abbiamo cercato di includere un mix eterogeneo di specializzazioni che ci permettesse di avere una più ampia gamma di punti di vista nei dati.

Per selezionare le pubblicazioni di diversi settori medici, abbiamo utilizzato le parole chiave "Chirurgia", "Anestesiologia", "Ginecologia", "Oftalmologia", "Medicina intensiva", "Neurologia", "Medicina di famiglia", "Cure primarie", "Radiologia", "Medicina nucleare", "Neuroradiologia", "Patologia", "Reumatologia", "Oncologia", "Dermatologia" insieme al connettore "AND" e alle parole chiave "Paesi Bassi", "Portogallo" e "Stati Uniti" in Google Scholar. Successivamente, attraverso tecniche di snowballing, sono stati identificati altri articoli e studiosi rilevanti.

Dato che questo studio mirava a sondare i medici, quando l'autore corrispondente di un articolo scientifico non era identificato come medico nella pubblicazione, abbiamo effettuato ulteriori ricerche su Google per confermare se lo studioso fosse effettivamente un medico. Ogni partecipante è stato quindi contattato attraverso le e-mail rese pubbliche nelle pubblicazioni scientifiche, in qualità di autore o coautore di una determinata pubblicazione e di medico.

Il gruppo finale di partecipanti a questo studio comprendeva medici (specializzandi e specialisti) di tredici diverse specialità, tra cui specialità mediche (medicina di famiglia, reumatologia, dermatologia, medicina intensiva, oncologia, neurologia), specialità chirurgiche (chirurgia, oftalmologia, ginecologia, anestesiologia, medicina riabilitativa, neurologia) e specialità diagnostiche (patologia, radiologia/medicina nucleare/neuroradiologia) con sede nei Paesi Bassi, in Portogallo e negli Stati Uniti.

Specializzazione	Portogallo	Paesi Bassi	STATI UNITI
Chirurgia	1	3	4
Anestesiologia	1	2	6
GINECOLOGIA	3	2	0
Oftalmologia	3	1	2
Medicina riabilitativa	1	1	1
Medicina intensiva	1	2	0
Neurologia	1	2	1
Medicina di famiglia	2	1	2
Radiologia/Medicina nucleare/Neuroradiologia	4	3	0
Patologia	1	6	0
Reumatologia	8	3	1
Oncologia	1	1	2
Dermatologia	1	3	0

Tabella 1. Partecipanti.

Un totale di 77 partecipanti ha completato con successo l'indagine, un numero adeguato per uno studio di q-metodologia con un q-set di 40 item [41,39]. Infatti, poiché la metodologia-Q mira solo a stabilire l'esistenza di particolari punti di vista, non è necessario un numero elevato di partecipanti. Inoltre, gli studi-

Q non richiedono un campione rigorosamente rappresentativo, ma piuttosto un campione di popolazione che contenga partecipanti con punti di vista rilevanti sull'argomento.

Siamo certi che il p-set di questo studio includa studiosi con punti di vista rilevanti sull'IA Health. Tuttavia, riconosciamo che, rivolgendoci a medici che hanno pubblicato di recente articoli scientifici, l'insieme dei partecipanti è composto principalmente da professionisti coinvolti in attività di ricerca e/o accademiche. Potremmo quindi non essere riusciti a rappresentare altre prospettive di operatori meno coinvolti nella ricerca.

In questo contesto, va anche chiarito che uno studio-Q non pretende che le dimensioni relative delle prospettive (in termini di numero di intervistati che vi aderiscono) riflettano la distribuzione della popolazione. In linea con il concetto che la metodologia q è una tecnica esplorativa piuttosto che confermativa, e riconoscendo il modo in cui è stato ottenuto il campione, ci asterremo dal trarre conclusioni quantitative sulle dimensioni delle prospettive e sulle differenze tra Paesi e specializzazioni. Le ricerche di conferma successive (ad esempio, volte a stabilire le opinioni della minoranza rispetto a quelle della maggioranza) dovrebbero basarsi su campioni rappresentativi.

3.5 STRUMENTO DI RACCOLTA DEL SONDAGGIO

I dati sono stati raccolti attraverso la versione html di FlashQ (Fig. 1, Fig. 2, Fig. 3),2 un software che permette di fare una cernita di q online. La distribuzione è stata codificata come una distribuzione a 11 punti che assomiglia a una distribuzione normale [-5, +5] con due celle poste sotto ciascuna coda (-5 e + 5), tre celle sotto entrambe le code -4 e 4, tre celle sotto entrambe le code -3 e 3, quattro celle sotto entrambe le code -2 e 2, cinque celle sotto entrambe le code -1 e 1, e sei celle sotto la coda 0. Ai partecipanti è stato chiesto di disporre le 40 affermazioni in base a una nozione soggettiva di disaccordo/accordo e successivamente è stato chiesto loro di fornire ulteriori commenti sulle affermazioni che avevano classificato a -5 e + 5. Ogni particolare disposizione delle affermazioni nella distribuzione forzata a cam-

pana è chiamata q-sort, quindi in questo studio i dati raccolti consistevano in 77 q-sort.

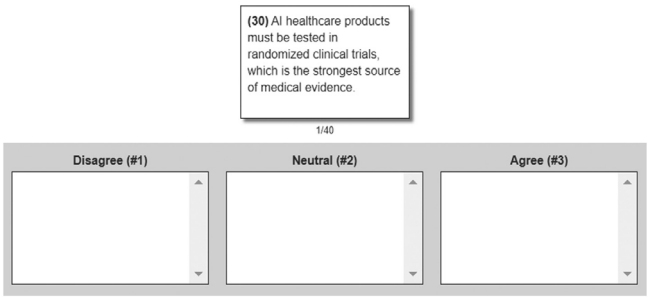


Fig. 1. Ordinamento dei contenitori: i partecipanti collocano le affermazioni randomizzate nei contenitori in disaccordo, neutro e d'accordo.

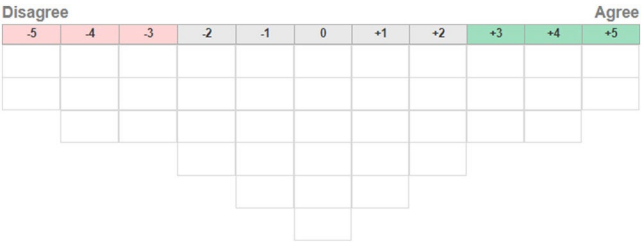


Fig. 2. Griglia di ordinamento: i partecipanti ordinano le affermazioni nella distribuzione a campana.

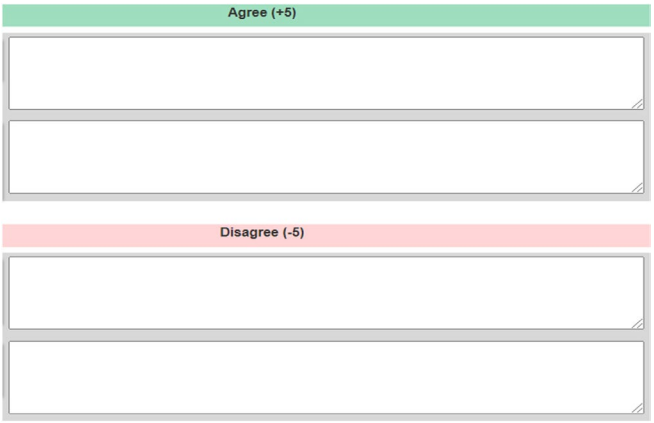


Fig. 3. Riquadro dei commenti: i partecipanti forniscono ulteriori commenti sulle affermazioni a cui hanno attribuito un punteggio di -5 e +5.

3.6 ANALISI

Come derivazione dell'analisi fattoriale, la metodologia q è una tecnica di riduzione dei dati che mira a ridurre un numero maggiore di variabili in un numero minore di fattori. Pertanto, il processo analitico della metodologia-Q si basa su tecniche di riduzione multivariata dei dati. Negli studi q, l'analisi dei dati comporta tre fasi principali: (i) estrazione dei fattori; (ii) rotazione dei fattori; (iii) interpretazione dei fattori. Per il processo analitico (fasi (i) e (ii)) abbiamo utilizzato PQMethod, un programma statistico che soddisfa i requisiti degli studi q3 [50].

La prima fase consiste nell'estrarre i fattori dalle q-sort raccolte in precedenza, riassumendo così tutte le risposte individuali in poche risposte rappresentative [44]. In questo studio, i fattori sono stati estratti attraverso l'analisi delle componenti principali (PCA), una tecnica di riduzione lineare che riduce la dimensionalità dei dati mantenendo la maggior parte della variazione nel set di dati, spesso utilizzata nell'analisi esplorativa dei dati [51]. L'identificazione di vettori ortogonali (componenti principali) lungo i quali la variazione è massima consente di ridurre i dati in poche componenti che rappresentano i modelli dominanti nei dati [52,51].

I fattori estratti sono stati successivamente ruotati per posizionare ciascun fattore in modo che il suo punto di vista si avvicini molto al punto di vista di un particolare gruppo di q-sort. Per la rotazione dei fattori, abbiamo utilizzato il Varimax, una rotazione ortogonale degli assi dei fattori che massimizza la varianza del carico di ciascun fattore rendendo più alti i carichi elevati e più bassi i carichi bassi. I Q-sort che hanno un carico elevato su un fattore avranno un carico basso su un altro, massimizzando così la distinzione e la differenziazione delle posizioni dei soggetti e minimizzando la correlazione tra i fattori [53]. Dopo aver ruotato un numero diverso di fattori e aver confrontato le distribuzioni dei tipi di definizione (segnalati automaticamente) tra i fattori, si è deciso di ruotare quattro fattori (Tabella 2). Questa soluzione

presenta il numero di fattori più alto ma interpretabile, in cui ogni fattore ha almeno tre tipi di definizione.

Cella vuota	Fattore 1	Fattore 2	Fattore 3	Fattore 4
Definizione delle specie	15	17	6	9
Autovalori	17.77	11.55	8.47	10.78
Varianza	17%	15%	11%	14%

Tabella 2. Panoramica dei fattori.

Ogni fattore è caratterizzato da un array di 40 punteggi (un punteggio per affermazione), che è un singolo q-sort configurato per rappresentare il punto di vista del fattore. Dato che i fattori hanno un numero diverso di serie definite, ogni punteggio nella matrice dei fattori è un punteggio standardizzato (z) per consentire il confronto tra fattori [54]. Gli array di fattori di ciascun fattore sono riportati nella Tabella 3.

Dichiarazione	Fattore 1	Fattore 2	Fattore 3	Fattore 4
1	0	-4	-5	-3
2	-1	-5	-3	-1
3	2	4	-3	2
4	-1	-4	-2	-4
5	1	-3	2	0
6	0	2	-1	1
7	0	0	-3	1
8	-1	1	1	0
9	4	5	4	4

10	-2	-5	-4	-2
11	-2	2	3	5
12	1	4	-1	2
13	-2	1	-5	-1
14	-2	-2	1	0
15	2	-1	1	-4
16	1	-2	-1	-2
17	-3	1	2	-2
18	-5	-1	0	-1
19	4	0	0	1
20	-5	-4	-1	1
21	-4	2	-2	-2
22	-4	-1	-4	-3
23	-1	1	-2	0
24	3	2	2	2
25	-4	-2	0	-5
26	0	3	4	4
27	3	0	1	2
28	1	-2	-1	0
29	0	3	1	0
30	3	5	5	-1
31	-1	-3	0	-1
32	4	3	5	4
33	5	4	3	5
34	0	0	0	-4
35	2	-1	3	1
36	5	0	2	3

37	2	1	4	3
38	-3	-3	-2	-5
39	-3	-1	-4	-3
40	1	0	0	3

Tabella 3. Quattro matrici di fattori in cui ogni matrice presenta i punteggi normalizzati [-5, +5] assegnati alle affermazioni del q-set dai partecipanti che hanno avuto un peso significativo sulla matrice di fattori.

Infine, l'ultima fase ha comportato l'analisi e l'interpretazione delle matrici di fattori delle quattro prospettive, al fine di comprendere le caratteristiche chiave di ciascuna prospettiva. A tal fine, abbiamo utilizzato il metodo crib sheet method [41]. Osservando le matrici di fattori, per ogni prospettiva abbiamo composto quattro categorie di base: (i) elementi con il ranking più alto nella matrice di fattori; (ii) elementi con il ranking più basso nella matrice di fattori; (iii) elementi classificati più in alto nel fattore *i* rispetto a qualsiasi altro fattore (di 2 o più unità); (iv) elementi classificati più in basso nel fattore *i* rispetto a qualsiasi altro fattore (di 2 o più unità). Si noti che l'interpretazione dei fattori fatta dagli autori è (intrinsecamente) soggettiva. È possibile che persone diverse arrivino a interpretazioni diverse dei quattro fattori sulla base degli stessi punteggi dei fattori. Tuttavia, dato che queste interpretazioni sono vincolate dai punteggi dei fattori, ci aspetteremmo che altri ricercatori arrivino a interpretazioni simili.

4. RISULTATI: PROSPETTIVE E INTERPRETAZIONI

In questo studio sono state identificate quattro diverse prospettive sull'IA sanitaria (Tabella 4). Le caratteristiche fondamentali di ciascuna prospettiva sono derivate dalle affermazioni classificate a -5 e + 5 [(*N* :| 5 |) dove *N* è il numero dell'affermazione e (|5|) può essere -5 o + 5], nonché dalle affermazioni classificate più alte o più basse rispetto alle matrici delle altre prospettive s [(N :| Pi |) dove *N* è il numero dell'affermazione, *P***i* è la pro-

spettiva con $i \in [1, 4]$ e |P**i| può essere -P**i o +P**i a seconda che l'affermazione sia classificata più bassa o più alta rispetto alle matrici delle altre prospettive]. Per illustrare ulteriormente ciascuna prospettiva, abbiamo incluso anche le dichiarazioni scritte dai partecipanti associate alle tipologie di definizione di ciascuna prospettiva, in merito alle affermazioni che hanno classificato come più alte e più basse.

P1	L'intelligenza artificiale è uno strumento utile: Lasciare che i medici facciano ciò per cui sono stati addestrati
P2	Le norme e i regolamenti sono fondamentali: Le aziende private pensano solo ai soldi
P3	L'etica è sufficiente: Ci si può fidare delle aziende private
P4	Strumenti di IA spiegabili: L'apprendimento è necessario e inevitabile

Tabella 4. Quattro prospettive sull'IA sanitaria.

4.1 PROSPETTIVA 1: L'IA È UNO STRUMENTO UTILE: LASCIATE CHE I MEDICI FACCIANO CIÒ PER CUI SONO STATI FORMATI

In questa prospettiva c'è una visione complessivamente positiva sull'implementazione della tecnologia AI nell'assistenza sanitaria. L'IA è considerata uno strumento utile che consentirà ai medici di avere il tempo necessario per concentrarsi su attività e competenze di alto livello (36:+5).

A sottolineare questa posizione, un partecipante ha scritto *Questo è l'obiettivo principale! Lasciare che i medici facciano ciò per cui sono stati formati – la medicina – e alleviare molti dei processi potenzialmente automatici e dispendiosi in termini di tempo che devono affrontare quotidianamente*. Un altro partecipante ha osservato che l'IA significa *meno tempo per il lavoro noioso e più tempo per il lavoro stimolante*. Un altro ancora ha riflettuto sui suoi primi giorni nel settore medico per sottolineare gli aspetti positivi dell'automazione *Come l'automazione dei test di laboratorio, l'IA libererà le mani e la mente degli operatori per concentrarsi su questioni di ordine superio-*

re. Come tirocinante, di notte dovevo girare i miei ematocriti. Non mi manca affatto!

Le argomentazioni tradizionali sollevate contro l'IA sanitaria sono sottovalutate in questa prospettiva. Non è problematico che l'IA sia una tecnologia a scatola nera, poiché i professionisti della salute utilizzano altre tecnologie a scatola nera, come la risonanza magnetica (16:1). C'è anche neutralità riguardo alla mancanza di conoscenza del contesto e di capacità di leggere gli indizi sociali (26:0). Su questa linea, un partecipante ha osservato che *il ruolo di un medico esperto è quello di prendere in considerazione ciò che una macchina o un'IA gli dice e fare il corretto collegamento con la realtà clinica*. Inoltre, questa prospettiva non condivide l'idea che l'IA peggiorerà i problemi dell'assistenza sanitaria, come i test eccessivi, le diagnosi eccessive e i trattamenti eccessivi (21:-4). Riguardo al ruolo dell'IA negli ospedali, un partecipante ha scritto che l'*IA giocherà un ruolo fondamentale nella stratificazione, assegnando così i pazienti a gruppi a basso rischio e ad alto rischio o a pazienti che rispondono a un determinato trattamento o a pazienti che non lo fanno. In questo modo si eviterà di sottoporre a test o trattamenti pazienti che non sono ritenuti efficaci.*

Nonostante le prospettive positive sull'IA, questa prospettiva sottolinea che i medici devono rimanere responsabili non solo nel processo decisionale medico (18:-5), ma anche partecipando al processo di progettazione della tecnologia (33:+5) (l'*IA aiuterà solo i medici a risolvere i loro dubbi clinici, ma l'ultima decisione non dovrebbe mai essere presa dall'IA*; *vedo l'IA come uno strumento aggiuntivo, non come qualcosa che sostituirà i medici o diminuirà l'autonomia e l'autorità. Un medico avrà sempre il verdetto finale*; *i progettisti di IA conoscono la tecnologia, ma hanno bisogno dei medici per progettare prodotti pertinenti*; *l'obiettivo di un determinato strumento di IA deve essere definito insieme ai medici per garantire la rilevanza clinica*; *poiché i medici non hanno competenze informatiche, gli ingegneri non hanno conoscenze mediche e necessità ospedaliere, i medici sono fondamentali nella progettazione dell'IA*).

Guardando al futuro dell'occupazione medica, secondo questa prospettiva gli ospedali basati sull'IA non dovrebbero rispar-

miare denaro impiegando medici meno qualificati (20:-5) (*con gli strumenti dell'IA che funzionano bene, la medicina progredirà verso un atto più preciso, con decisioni basate sull'opinione di un team multidisciplinare, per cui saranno necessari soprattutto medici altamente qualificati*), ma si ritiene che, anche se l'IA non sostituirà i medici, i medici che la usano sostituiranno quelli che non la usano (19-4) *Come in ogni settore del progresso tecnico, l'IA è uno strumento che sarà abbracciato da chi è all'avanguardia. Coloro che non lo faranno, come i chirurghi che non hanno mai imparato la laparoscopia, vedranno diminuire il loro* campo di *applicazione.*

4.2 PROSPETTIVA 2: NORME E REGOLAMENTI SONO FONDAMENTALI: LE AZIENDE PRIVATE PENSANO SOLO AI SOLDI

Nella seconda prospettiva identificata in questo studio, vi è una chiara visione negativa della tecnologia AI (21:2; 31-3) e una chiara sfiducia nelle aziende sanitarie private. Si ritiene che l'industria tecnologica non sia ben allineata con i valori fondamentali della sanità (le *aziende private pensano solo ai SOLDI*), abbia una scarsa conoscenza della medicina (13:1) e rappresenti un rischio di comportamento monopolistico (12:4).

Per quanto riguarda il rischio di un comportamento monopolistico, c'è una grande preoccupazione per le implicazioni del possesso dei dati medici da parte di queste grandi aziende. La privacy è annunciata come un valore etico fondamentale in campo medico, anche in un futuro in cui l'assistenza sanitaria si affida fortemente all'IA (2:-5; 4:-4). In particolare, si teme per le implicazioni sul rapporto medico-paziente. Un partecipante ha osservato che *quando un paziente teme che la riservatezza venga violata nei confronti di un'azienda tecnologica o di una compagnia assicurativa, potrebbe non fornire informazioni complete o evitare di sottoporsi a un trattamento, con conseguente aggravamento della malattia.* Ci sono anche preoccupazioni riguardo al potere che deriverebbe dal possesso di tali dati. Un partecipante ha messo in guardia sul fatto che le *aziende di IA sanitaria avrebbero troppo potere se queste informazioni non fossero anonime. Venderebbero informazioni su perso-*

ne specifiche alle aziende farmaceutiche, alle società di assunzione, alle compagnie di assicurazione…

Poiché in questa prospettiva non ci si può fidare delle aziende private, si pone una forte enfasi sulle norme e sui regolamenti per tenere sotto controllo queste società. Si ritiene che il mantra tecnologico "fail fast and fix later" [sbaglia adesso e correggi dopo] metta a rischio i pazienti e che le autorità di regolamentazione non stiano facendo abbastanza per mantenere i consumatori al sicuro (29:3). Pertanto, anche se le tecnologie di IA per la salute e i loro sviluppatori dovrebbero essere vincolati da un'etica medica di base (9:5), è necessario stabilire regole chiare in materia di responsabilità, dati e certificazione dei prodotti.

Inoltre, si ritiene che le aziende tecnologiche debbano essere responsabili degli errori medici anche se tale responsabilità ostacola il progresso tecnologico (10:-5). Un partecipante ha affermato che il *profitto richiede un rischio e le aziende devono sopportarlo. Gli sviluppatori devono essere responsabili se vogliono entrare nell'esigente arena delle cure*. Queste tecnologie dovrebbero essere utilizzate nell'assistenza sanitaria solo quando l'IA sarà dotata di norme e regolamenti chiari sull'uso, l'archiviazione e l'anonimizzazione dei dati (3:4).

Sulla stessa linea, si sottolinea anche che i prodotti sanitari di IA devono essere testati in studi clinici randomizzati, che sono la fonte più solida di prove mediche (30: +5). Un partecipante ha illustrato chiaramente questo punto affermando che, *poiché gli strumenti di diagnosi e cura dell'IA influiscono direttamente sulla salute dei pazienti, dovrebbero essere sottoposti agli standard più elevati, come avviene normalmente in medicina per le nuove strategie di diagnosi e cura. Non vedo perché questo dovrebbe essere diverso per l'IA rispetto ai nuovi "test diagnostici convenzionali" o ai nuovi farmaci*.

4.3 PROSPETTIVA 3: *L'ETICA È SUFFICIENTE: LE AZIENDE PRIVATE POSSONO ESSERE AFFIDABILI*

La caratteristica più evidente di questa prospettiva è la visione complessivamente positiva delle aziende di IA. Secondo

questo punto di vista, non è indesiderabile che queste aziende inizino a operare nello spazio sanitario (13:-5). Inoltre, non ci sono grandi preoccupazioni riguardo al rischio di comportamenti monopolistici (12:-1).

Una potenziale spiegazione di questo giudizio positivo sulla tecnologia è la percezione che gli attuali sistemi sanitari facciano già molto affidamento sulla tecnologia e sulle aziende tecnologiche. Come ha osservato un partecipante, il potenziale dell'automazione negli ospedali è *ridicolo. L'automazione funziona già negli ospedali*.

Piuttosto che soffermarsi su norme e regolamenti (3:-3), l'etica di per sé è sufficiente a fondare il settore privato. La tecnologia di IA deve essere allineata con i principi bioetici (32:+5), come la privacy (1:-3; 2:-3), che dovrebbe rimanere un valore medico fondamentale (4:-2). Nonostante la fiducia nelle aziende tecnologiche, anche in questa prospettiva si sottolinea la necessità di testare i prodotti sanitari di IA (30:5).

Secondo questa prospettiva, l'IA non aumenterà la discriminazione basata su problemi medici futuri previsti (7:-3) e quindi il miglioramento dell'equità e dell'inclusione non deve essere la priorità principale nello sviluppo e nell'impiego dell'IA nell'assistenza sanitaria (6:-1).

4.4 PROSPETTIVA 4: *STRUMENTI DI IA SPIEGABILI: L'APPRENDIMENTO È NECESSARIO E INEVITABILE*

La spiegabilità è un valore chiave in questa prospettiva. Per cogliere i benefici dell'IA, i medici devono comprendere e guidare il progresso tecnologico dell'IA. Un partecipante ha scritto che l'IA *non dovrebbe mai essere una "scatola nera". I medici dovrebbero essere in grado di spiegare i risultati degli strumenti di IA con un ragionamento*. Non solo gli operatori sanitari devono sapere come funzionano gli strumenti medici di IA (15:-5), ma di fatto tali strumenti dovrebbero essere utilizzati solo se i medici capiscono come vengono prese le decisioni di IA (11:5). A questo proposi-

to, un partecipante ha osservato che l'*adozione dell'IA migliorerà se i medici capiranno la "scatola nera"*.

Secondo questa prospettiva, i medici sono interessati a conoscere l'IA e l'informatica (25:+5) e hanno il tempo di imparare a usare dispositivi medici complessi basati sull'IA (3:-4). Un partecipante ha osservato che, *in generale, l'intelletto dei medici è sottovalutato e poco valutato dai tecnici*; un altro ha sottolineato che *i medici non possono lavorare senza computer e li usano quotidianamente per la registrazione. L'apprendimento è necessario e inevitabile.*

Si ritiene problematico che le macchine manchino di conoscenza del contesto e di capacità di leggere gli indizi sociali (26:4) e che sia difficile rendere operativa la pratica clinica per una macchina (38:-5). Pertanto, i medici devono partecipare al processo di progettazione dell'IA per l'assistenza sanitaria (33:5). Di conseguenza, un partecipante ha scritto che l'*IA è qui per restare (credo), e i medici sono i più adatti a regolare e migliorare i vari algoritmi* ecc. *attualmente in fase di progettazione*.

5. DISCUSSIONE

Le prospettive identificate in questo studio rivelano punti di vista diversi e spesso contraddittori sull'IA sanitaria. La comprensione di questi valori e tensioni sottostanti è importante per rendere operative le questioni etiche associate all'implementazione delle tecnologie di IA in ambito sanitario. In ultima analisi, ci si aspetta che tale operazionalizzazione porti a dibattiti e politiche più significativi verso una diffusione eticamente allineata dell'IA sanitaria.

Il nostro studio offre un'analisi sistematica delle prospettive dei medici sull'IA sanitaria. È possibile osservare elementi delle quattro prospettive riportate nella letteratura attuale. Diversi articoli hanno riportato risultati che riconducono alla P1 (L'IA è uno strumento utile: lasciate che i medici facciano ciò per cui sono stati formati) per quanto riguarda gli atteggiamenti positivi sull'uso dell'IA come tecnologia di supporto [18,19,26,27,25,24].

Anche la necessità di ampliare e migliorare la formazione medica in materia di IA è ben affrontata in letteratura [18,20,21,26,25] e si ricollega a P4 (Strumenti di IA spiegabili: l'apprendimento è necessario e inevitabile), che sostiene che i medici devono comprendere l'IA. Inoltre, nessuna delle prospettive identificate in questo studio rivela preoccupazioni relative alla sostituzione del lavoro, un dato in linea con gli studi riportati in letteratura [18,26,17,32,24]. La forza del nostro lavoro risiede nel fatto che, poiché i partecipanti sono stati intervistati su un'ampia gamma di questioni etiche relative all'IA sanitaria, le prospettive emerse forniscono un quadro più completo delle opinioni morali degli operatori.

Ciascuna prospettiva fornisce spunti di riflessione sulle prospettive dell'etica dell'IA in campo sanitario e anche su particolari questioni etiche, come l'equità, la spiegabilità e la progettazione etica, che devono essere prese in considerazione nei processi di implementazione delle tecnologie IA in campo sanitario.

Per quanto riguarda le prospettive sull'etica della salute dell'IA, P1 e P2 (Norme e regolamenti sono fondamentali: Le aziende private pensano solo ai soldi) rappresentano opinioni piuttosto convenzionali sulla tecnologia dell'IA, in contrasto con P3 (L'etica è sufficiente: le aziende private possono essere affidabili).

La percezione che gli strumenti medici basati sull'IA miglioreranno l'efficienza in ambito clinico si basa su decenni di successi nello sviluppo di sofisticate tecnologie mediche. In generale, questa prospettiva (P1) è allineata con le narrazioni presentate dalle aziende tecnologiche, che tendono a concentrarsi sui benefici della tecnologia e dell'automazione per l'assunzione di compiti ripetitivi. Racconti simili sono presentati dagli sviluppatori di altre tecnologie basate sull'IA, come il veicolo autonomo [55].

Proiettando le tecnologie mediche basate sull'IA come un altro tipo di strumento medico, le conversazioni di ordine superiore sulle questioni etiche associate a questa tecnologia, come l'equità o l'autonomia umana, vengono in qualche modo evitate. Tuttavia, dato il potenziale dirompente di queste tecnologie, sono necessarie ulteriori riflessioni sull'etica. Le aziende tecno-

logiche e gli sviluppatori dovrebbero riconoscere la singolarità dell'IA e garantire che il processo di progettazione sia guidato da considerazioni etiche.

Quando l'enfasi non è solo sui benefici della tecnologia AI, la regolamentazione è spesso vista come la soluzione per garantire la sicurezza dei consumatori. I medici che sono esitanti nei confronti degli strumenti medici basati sull'IA, ritengono che le norme e i regolamenti siano un elemento cruciale nella transizione verso un'assistenza sanitaria basata sull'IA. Questa prospettiva è ben allineata con la tradizione nei settori dell'assistenza sanitaria e della medicina, che sono notoriamente fortemente regolamentati [56]. La regolamentazione dei dispositivi medici basati sull'IA (noti anche come Software come dispositivo medico basato sull'intelligenza artificiale/apprendimento automatico) presenta sfide uniche ed è nota per essere in ritardo rispetto alla tecnologia [57].

Al contrario, una prospettiva meno convenzionale rispetto all'etica dell'AI Health riportata in questo studio (P3: L'*etica è sufficiente: ci si può fidare delle aziende private*) ritiene che non sia necessaria una pesante regolamentazione delle aziende tecnologiche. Rifiutandosi di demonizzare le aziende tecnologiche di IA, questa prospettiva fornisce una visione positiva un po' insolita di questi attori della salute. Si basa sull'idea che la consapevolezza etica sia sufficiente per ottenere la fiducia delle aziende. Recentemente è stato sostenuto che le aziende di IA dovrebbero promuovere l'etica della virtù, piuttosto che le tradizionali linee guida infuse di deontologia, come forma efficace per garantire un comportamento etico nelle aziende [10].

Anche se rinunciare completamente alle regolamentazioni può essere irrealistico, le aziende tecnologiche private dovrebbero comprendere che per essere accettate come attori etici nello spazio sanitario, pur puntando ai profitti, devono promuovere ambienti e pratiche etiche.

In merito a particolari questioni etiche, il nostro studio ha messo in luce tensioni e punti di vista contraddittori che dovrebbero essere presi in considerazione nei dibattiti futuri, in

particolare per quanto riguarda l'equità, la spiegabilità e la progettazione etica.

Equità, non discriminazione e giustizia si riferiscono alla ragionevolezza e all'imparzialità delle azioni. Le tecnologie basate sull'IA sono progettate e prodotte dall'uomo e si basano ampiamente sui dati, quindi sono esposte a errori, giudizi errati e pregiudizi che possono entrare nel ciclo di vita dell'innovazione e creare pregiudizi [58,59]. Ci sono diverse preoccupazioni riguardo a risultati distorti o discriminatori nel contesto dell'IA per la salute. Un dispositivo medico distorto opera in modo tale da produrre svantaggi per alcuni gruppi demografici e influenzare le disuguaglianze nella salute [60]. Ai dispositivi medici alimentati dall'IA sono associati diversi tipi di pregiudizi: pregiudizi fisici (il design del dispositivo medico svantaggia alcuni gruppi demografici in base a tratti fisici come il colore della pelle), pregiudizi computazionali (i set di dati di addestramento che servono come input al dispositivo medico non sono rappresentativi della popolazione) e pregiudizi interpretativi (il dispositivo medico è soggetto a un'inferenza distorta delle letture) [60].

Il nostro studio indica che ci sono più preoccupazioni su questi temi in P2 e meno in P3, il che non sorprende viste le restanti caratteristiche di queste prospettive. Tuttavia, in generale, tutte le prospettive sono abbastanza neutrali quando si parla di equità e discriminazione nell'IA sanitaria. Questa neutralità potrebbe essere un difetto di questo studio, in particolare per quanto riguarda le affermazioni selezionate sull'equità, ma potrebbe anche indicare che i medici sono poco informati su questi temi o semplicemente non li considerano rilevanti o urgenti. Recentemente sono stati riportati in letteratura alcuni casi concreti di dispositivi medici non equi (ad esempio, è stato riportato che i pulsossimetri non sono altrettanto accurati nel misurare l'ossigenazione del sangue nei pazienti di colore [60]), ma è possibile che questi argomenti rimangano in gran parte astratti per la maggior parte dei medici. Una futura ricerca empirica dovrebbe esplorare ulteriormente le opinioni degli operatori su questioni di equità e pregiudizi.

Un'altra questione importante associata all'etica dell'IA sanitaria è la spiegabilità. Un modello spiegabile fornisce informazioni *interpretabili* (descrizione di un sistema in modo comprensibile per l'uomo) e *complete* (descrizione accurata del funzionamento di un sistema) sul sistema [61]. La sfida della spiegabilità consiste quindi nel raggiungere sia l'interpretabilità che la completezza, dato che le spiegazioni accurate non sono facilmente interpretabili e queste ultime spesso mancano di potere predittivo [61]. Le tecnologie mediche alimentate dall'IA si basano su algoritmi complessi che non sono facilmente interpretabili, quindi noti come black-box.

Il nostro studio ha rivelato punti di vista contrastanti rispetto alle tecnologie di IA per la salute spiegabili. Secondo P1, la mancanza di spiegabilità nei dispositivi alimentati dall'IA non è un problema, poiché i professionisti della salute hanno utilizzato altre tecnologie complesse, come la risonanza magnetica, che assomigliano anch'esse a scatole nere. Al contrario, P4 ritiene che la spiegabilità sia un valore chiave e che, per trarre i benefici dell'IA, i medici debbano comprendere le complessità dei dispositivi medici alimentati dall'IA.

Il confronto con la risonanza magnetica è spesso richiesto in questa letteratura, ma non è ampiamente accettato. In effetti, la risonanza magnetica è una tecnologia medica complessa e non ci si aspetta che gli operatori conoscano la fisica e la matematica alla base di questa tecnologia. Tuttavia, gli algoritmi che fanno funzionare questi sistemi sono effettivamente spiegabili e comprensibili per gli sviluppatori. Diversamente, la sfida della spiegabilità associata all'IA non è contingente ai medici, ma anche agli sviluppatori in generale. La mancanza di conoscenza delle regole decisionali che sostengono un determinato risultato è particolarmente problematica in ambito sanitario. Come ha osservato un partecipante, *un buon operatore sanitario non si affiderà mai ciecamente a una singola misura senza conoscere la storia del paziente.*

I punti di vista contrastanti riportati nel nostro studio supportano la necessità di ulteriori ricerche empiriche per determinare se i professionisti che condividono prospettive diverse

rispetto alla spiegabilità interagiscano in modo diverso con lo stesso algoritmo [62].

Entrambi i valori di equità e spiegabilità esplorati sopra dovrebbero essere presi in considerazione nella progettazione etica dell'IA sanitaria. Il nostro studio mostra chiaramente che, a prescindere dalle prospettive positive o negative del settore, tutte le prospettive ritengono che i medici debbano partecipare al processo di progettazione delle tecnologie sanitarie di IA.

Approfondendo i commenti dei partecipanti, sembra che i medici ritengano che il successo e la rilevanza clinica dell'IA dipendano dal coinvolgimento dei professionisti nella progettazione e nello sviluppo della tecnologia. Le ragioni addotte dai partecipanti vanno oltre le conoscenze mediche (*i medici non hanno competenze informatiche, gli ingegneri non hanno conoscenze mediche ed esigenze ospedaliere, quindi i medici sono fondamentali nella progettazione dell'IA.*), e comprendono anche il ragionamento clinico (*senza sapere come pensa un clinico, l'IA non sarebbe uno strumento utile*) e il ruolo sociale dei medici (*i medici sono addestrati e dedicati al processo decisionale etico e sociale. Sono costruttori naturali di ponti tra una realtà medica/tecnica complessa e lo spazio personale di un singolo paziente. L'attraversamento di questo ponte è fondamentale perché ogni nuovo sviluppo abbia una ragione di esistere. Senza il coinvolgimento dei medici, le principali parti interessate allo sviluppo vengono trascurate*).

Lo sviluppo della tecnologia IA sanitaria dovrebbe essere uno sforzo multidisciplinare. È ancora da vedere se i professionisti agiranno come consulenti o, come ammonito in P4, assumeranno un ruolo più importante nello sviluppo della tecnologia. Ulteriori ricerche dovrebbero esplorare modelli di sviluppo tecnologico in grado di integrare gli elementi sopra menzionati nel processo di progettazione.

Il nostro studio esplorativo ha rivelato quattro prospettive sull'IA sanitaria, che riteniamo possano contribuire a dare forma a futuri dibattiti e processi di progettazione etica. Esistono opinioni contrastanti sull'etica dell'IA sanitaria in generale, ma anche su questioni etiche particolari come la spiegabilità. Si osserva che i medici sono più preoccupati del ruolo delle grandi

aziende nell'assistenza sanitaria e meno consapevoli o preoccupati di questioni di livello superiore e spesso astratte come l'equità, i pregiudizi e le disuguaglianze sanitarie.

Questo studio presenta importanti limitazioni. La prima limitazione è legata al processo di filtraggio, in cui l'insieme di 353 affermazioni recuperate dalla letteratura scientifica e divulgativa è stato ridotto a 40 affermazioni. Questo processo è stato effettuato dagli autori, che non hanno una formazione medica. Riconosciamo che il contributo di un medico in questo processo di filtraggio avrebbe aggiunto valore a questo studio, permettendoci di comprendere meglio la rilevanza e la conoscenza dei medici sugli argomenti trattati nelle dichiarazioni. Un'altra limitazione è legata al reclutamento dei partecipanti. Come già accennato, la maggior parte dei partecipanti a questo studio è stata reclutata attraverso pubblicazioni scientifiche recenti. Potremmo quindi non aver colto le prospettive dei medici che sono meno coinvolti nelle attività accademiche e di ricerca. Reclutando medici da tre Paesi occidentali, inoltre, non abbiamo incluso nel nostro p-set medici provenienti da Paesi in via di sviluppo, che avrebbero potuto contribuire con ulteriori prospettive sull'IA sanitaria. Infine, mentre in questo studio ci siamo concentrati sulle prospettive relative alle applicazioni dell'IA nell'assistenza sanitaria in generale, è necessario ampliare la letteratura e analizzare domini e compiti particolari dell'IA sanitaria.

6. CONCLUSIONE

Affinché l'IA possa esprimere il suo potenziale nel complesso settore della sanità, è necessario tenere conto dell'etica. In questo studio empirico abbiamo intervistato medici con sede nei Paesi Bassi, in Portogallo e negli Stati Uniti su un'ampia gamma di questioni etiche relative all'IA sanitaria. L'indagine ci ha permesso di individuare diverse prospettive sull'etica che circonda l'impiego dell'IA sanitaria.

Abbiamo identificato quattro prospettive principali: P1: l'*intelligenza artificiale è uno strumento utile: Lasciare che i medici facciano ciò per cui sono stati formati*; P2: *Regole e regolamenti sono fondamentali: Le aziende private pensano solo ai soldi!*; P3: L'*etica è sufficiente: Le aziende private possono essere affidabili*; e P4: *Strumenti di IA spiegabili: L'apprendimento è necessario e inevitabile!*

Ciascuna prospettiva fornisce indicazioni preziose sulle questioni etiche che dovrebbero essere rese operative e tenute in considerazione nella progettazione e nello sviluppo di queste tecnologie. Il nostro studio rivela punti di vista contrastanti sull'etica associata all'IA sanitaria. Si osserva inoltre che i medici sono per lo più preoccupati del ruolo delle grandi aziende nell'assistenza sanitaria e meno consapevoli o preoccupati di questioni di livello superiore come l'equità, i pregiudizi e le disuguaglianze sanitarie. A prescindere dalle prospettive positive e negative del settore, il nostro studio ha rivelato che i medici ritengono di dover partecipare al processo di progettazione. Questi risultati sono utili punti di partenza per una discussione proficua tra professionisti del settore medico, stakeholder dell'industria e responsabili politici.

Data la natura esplorativa di questa ricerca, ci sono ampie possibilità di confermare le direzioni di ricerca e di esplorare come tradurre queste prospettive in intuizioni attuabili e modelli di progettazione per i diversi attori della salute.

6 - INTELLIGENZA ARTIFICIALE SPIEGABILE (XAI) NELL'ANALISI DELLE IMMAGINI MEDICHE BASATA SUL DEEP LEARNING

Tratto e tradotto da

Bas H.M.van der Velden, Hugo J.Kuijf, Kenneth G.A. Gilhuijs, Max A.Viergever (2022), Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. In (Ed.) Medical Image Analysis – Volume 79, luglio 2022, 102470.

https://doi.org/10.1016/j.media.2022.102470

Le parti omesse dal curatore rispetto all'originale sono indicate dal segno **[...]**

HIGHLIGHTS

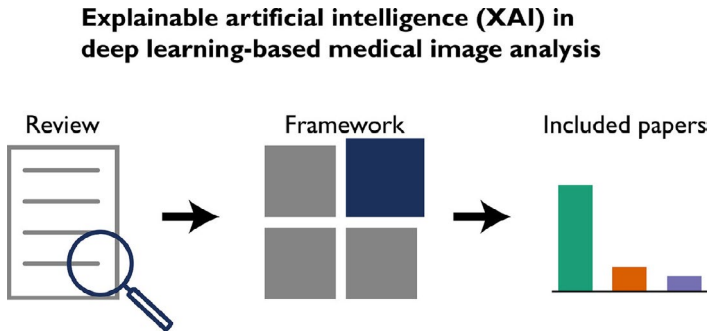
- Questo documento esamina oltre 200 articoli che utilizzano l'intelligenza artificiale spiegabile (XAI) nell'analisi delle immagini mediche basata sul deep learning.
- I papers esaminati sono stati classificati in base a un framework XAI.
- Vengono identificate le tendenze e le prospettive future della XAI nell'analisi delle immagini mediche.

ABSTRACT

Con l'aumento dei metodi basati sull'apprendimento profondo, cresce la richiesta di spiegabilità di tali metodi, soprattutto in aree decisionali ad alto rischio come l'analisi delle immagini mediche. Questa ricerca presenta una panoramica dell'intelligenza

artificiale spiegabile (XAI, Explainable Artificial Intelligence) utilizzata nell'analisi delle immagini mediche basata sul deep learning. Viene introdotto un quadro di criteri XAI per classificare i metodi di analisi delle immagini mediche basati sul deep learning. Gli articoli sulle tecniche XAI nell'analisi delle immagini mediche vengono quindi esaminati e classificati in base al quadro di riferimento e alla posizione anatomica. L'articolo si conclude con una panoramica delle opportunità future per la XAI nell'analisi delle immagini mediche.

ABSTRACT GRAFICO



1. INTRODUZIONE

L'apprendimento profondo ha fatto compiere enormi progressi all'analisi automatizzata delle immagini. In precedenza, l'analisi delle immagini veniva comunemente eseguita utilizzando sistemi completamente progettati da esperti di dominio umani. Ad esempio, un tale sistema di analisi delle immagini poteva consistere in un classificatore statistico che utilizzava proprietà di un'immagine create a mano (cioè le caratteristiche) per eseguire un determinato compito. Le caratteristiche comprendono proprietà dell'immagine di basso livello, come i bordi o gli angoli, ma anche proprietà dell'immagine di livello superiore, come

il bordo spaccato di un tumore. Nell'apprendimento profondo [deep learning], queste caratteristiche vengono apprese da una rete neurale (a differenza di quelle create a mano) per fornire un risultato (o output) ottimale in base a un input. Un esempio di sistema di deep learning potrebbe essere l'output "cancro" dato l'input di un'immagine che mostra un cancro.

Le reti neurali sono tipicamente composte da molti strati collegati tra loro da numerose relazioni non lineari. Anche se si dovessero ispezionare tutti questi strati e descrivere le loro relazioni, non è possibile comprendere appieno come la rete neurale sia giunta alla sua decisione. Pertanto, il deep learning è spesso considerato una "scatola nera". In diversi campi di applicazione si sta diffondendo la preoccupazione che queste scatole nere possano essere in qualche modo distorte e che tale distorsione passi inosservata. Soprattutto nelle applicazioni mediche, ciò può avere conseguenze di vasta portata.

È stata lanciata una richiesta di approcci per comprendere meglio la scatola nera. Tali approcci sono comunemente indicati come deep learning interpretabile o intelligenza artificiale spiegabile (XAI) (Adadi e Berrada, 2018; Murdoch et al., 2019). Questi termini vengono comunemente scambiati; noi utilizzeremo il termine XAI. Alcune iniziative XAI degne di nota sono quelle della Defense Advanced Research Projects Agency (DARPA) degli Stati Uniti e le conferenze su equità, responsabilità e trasparenza dell'Association for Computing Machinery (ACM FAccT).

La posta in gioco nel processo decisionale medico è spesso alta. Non sorprende che gli esperti di medicina abbiano espresso la loro preoccupazione per la natura "black box" del deep learning (Jia et al., 2020), che rappresenta l'attuale stato dell'arte nell'analisi delle immagini mediche (Litjens et al., 2017; Meijering, 2020; Shen et al., 2017). Inoltre, normative come il Regolamento generale sulla protezione dei dati (GDPR, articolo 15) dell'Unione Europea richiedono il diritto dei pazienti di ricevere informazioni significative su come è stata presa una decisione.

I ricercatori nel campo dell'imaging medico utilizzano sempre più spesso XAI per spiegare i risultati dei loro algoritmi. Una

spiegazione può essere considerata valida se fornisce indicazioni su come una rete neurale è giunta alla sua decisione e/o se è in grado di rendere la decisione comprensibile. In questa indagine, ci proponiamo di fornire una panoramica completa degli articoli che utilizzano XAI nell'analisi delle immagini mediche. Abbiamo scelto di concentrarci esclusivamente sui lavori che hanno utilizzato XAI basato sul deep learning nell'analisi delle immagini mediche.

La strategia di ricerca per l'inclusione degli articoli è stata la seguente: Abbiamo utilizzato la query di ricerca "(explainable deep learning OR interpretable deep learning OR XAI OR interpretable machine learning OR explainable machine learning) AND (medical imaging OR medical image analysis)" in SCOPUS. Abbiamo incluso articoli provenienti da riviste e conferenze peer reviewed. Abbiamo analizzato i risultati delle query utilizzando il toolbox Active learning for Systematic Reviews (van de Schoot et al., 2021). Questo toolbox utilizza l'apprendimento attivo per ordinare gli articoli da quelli più rilevanti a quelli meno rilevanti, aggiornandoli in base agli input dell'utente. Inoltre, abbiamo discusso con i colleghi e abbiamo utilizzato un approccio di tipo "snowballing", indagando sui lavori a cui fanno riferimento i lavori inclusi e sui lavori che fanno riferimento ai lavori inclusi. Abbiamo letto il titolo e l'abstract di ciascuno di questi articoli e abbiamo sfogliato il contenuto del documento se non eravamo sicuri di includerlo. In caso di pubblicazioni multiple degli stessi autori sullo stesso argomento, abbiamo scelto la pubblicazione su rivista o la pubblicazione più recente in caso di pubblicazioni multiple a conferenze. I lavori fino a ottobre 2020 sono stati inclusi nell'indagine.

L'indagine è strutturata come segue: Nella Sezione 2, introduciamo la tassonomia di XAI e descriviamo un framework per classificare le tecniche XAI. Nella Sezione 3, i lavori discussi sono caratterizzati in base a questo framework XAI. Discuteremo le applicazioni delle tecniche XAI nell'analisi delle immagini mediche. Nel caso di più articoli che utilizzano la stessa tecnica, discuteremo alcuni dei primi adottati e riassumeremo il resto de-

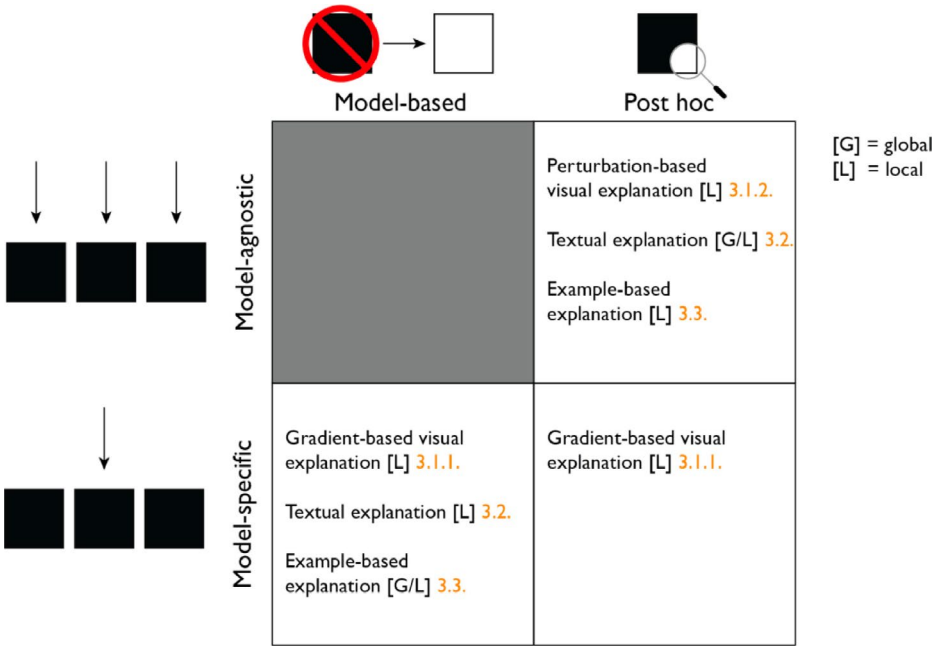
gli articoli nelle tabelle. Dato che le tecniche XAI spesso provengono dalla computer vision, approfondiremo i lavori che hanno adattato le tecniche XAI dalla computer vision aggiungendo le conoscenze di dominio del campo dell'imaging medico. I lavori sono raggruppati nelle tabelle in base al metodo di spiegazione e alla posizione anatomica. Questa rassegna si aggiunge a quella di Reyes et al. (2020); poiché questi ultimi hanno discusso principalmente tecniche di computer vision, senza valutare estensivamente l'adattamento di tali tecniche all'analisi di immagini mediche. Inoltre, descriviamo se e come le tecniche di computer vision sono state adattate specificamente per l'analisi delle immagini mediche. Questa rassegna si aggiunge a quella di Huff et al. (2021), in quanto questi ultimi si sono concentrati principalmente su esempi di spiegazione visiva, mentre la nostra rassegna mira a un approccio più olistico che include spiegazioni non visive, critiche sulla XAI e metodi di valutazione della XAI. Inoltre, abbiamo effettuato un'indagine sistematica dei papers che riflettono lo stato attuale del campo della XAI nell'imaging medico. Nella Sezione 4, discutiamo i pro e i contro delle tecniche XAI discusse. La Sezione 5 conclude l'indagine discutendo lo stato dell'arte della XAI nell'analisi delle immagini mediche e una prospettiva delle opportunità della XAI.

2. STRUTTURA DELL'INTELLIGENZA ARTIFICIALE SPIEGABILE (XAI)

In questa sezione, forniremo una breve panoramica delle tecniche di Explainable Artificial Intelligence (XAI) presenti nel deep learning per l'analisi delle immagini mediche. Per indagini esaustive incentrate esclusivamente sulla XAI, si rimanda ad Adadi e Berrada (2018) e Murdoch et al. (2019).

Distingueremo le tecniche XAI in base a tre criteri: basate sul modello rispetto a quelle post hoc, specifiche rispetto a quelle agnostiche e globali rispetto a quelle locali (cioè, l'ambito della spiegazione). Il quadro di questi tre criteri è adattato dalle inda-

gini di Adadi e Berrada (2018) e Murdoch et al. (2019) ed è raffigurato nella Fig. 1. I paragrafi seguenti descrivono questi criteri.



Download: Scarica l'immagine ad alta risoluzione (291KB) Download: Scarica l'immagine a grandezza naturale Fig. 1. *Il quadro dell'Intelligenza Artificiale Esplicita (XAI) proposto in questo lavoro. Una panoramica approssimativa delle tecniche XAI (discusse nella Sezione 3) è classificata secondo questo schema. Il numero arancione si riferisce al numero della sezione del manoscritto in cui è descritta la tecnica XAI.*

2.1 SPIEGAZIONE BASATA SUL MODELLO E SPIEGAZIONE POST HOC

La prima distinzione che facciamo è tra spiegazione basata sul modello e spiegazione post hoc (Fig. 1).

2.1.1 Spiegazione basata sul modello

La spiegazione basata su modelli si riferisce a modelli, ad esempio un modello di regressione lineare o una macchina vettoriale di supporto, che sono abbastanza semplici da essere com-

presi, ma abbastanza sofisticati da adattarsi bene a una relazione tra input e output (Murdoch et al., 2019). Questi sono spesso i modelli tradizionali di apprendimento automatico. Esempi di spiegazioni basate su modelli impongono l'uso di una quantità limitata di caratteristiche (cioè la sparsità) o impongono che un essere umano sia in grado di ragionare internamente sull'intero processo decisionale del modello (cioè la simulabilità) (Murdoch et al., 2019). Ad esempio, i modelli che impongono la sparsità, come il least absolute shrinkage and selection operator [Operatore di selezione e restringimento minimo assoluto] (LASSO, Tibshirani (1996), costringono molti coefficienti a zero. Quindi, un sottoinsieme selezionato di caratteristiche porta a un output, rendendo spiegabile il costrutto interno di questo modello.

Poiché la nostra indagine si concentra sui metodi XAI per l'apprendimento profondo, la spiegazione basata sul modello, che impone la sparsità o la simulabilità, non è fattibile. L'apprendimento profondo utilizza una rete neurale profonda, in genere con migliaia o milioni di pesi, che non è né sparsa né adatta a un essere umano per simulare internamente e ragionare sull'intero processo decisionale dei modelli. Tuttavia, uno dei metodi menzionati da Murdoch et al. (2019) è l'ingegneria delle caratteristiche basata sui modelli, ossia approcci automatizzati per la costruzione di caratteristiche spiegabili.

2.1.2 Spiegazione post hoc

L'analisi di un modello addestrato (ad esempio, una rete neurale nel deep learning) per ottenere una comprensione delle relazioni apprese viene definita spiegazione post hoc. Una distinzione importante tra la spiegazione post hoc e la spiegazione basata sul modello è che la prima addestra una rete neurale e successivamente tenta di spiegare il comportamento della rete risultante dalla scatola nera, mentre la seconda costringe la rete neurale a essere spiegabile.

I metodi che forniscono spiegazioni post hoc includono l'ispezione delle caratteristiche apprese, l'importanza delle carat-

teristiche e l'interazione delle caratteristiche (Abbasi-Asl e Yu, 2017; Olden et al., 2004; Tsang et al. 2018; nonché la spiegazione visiva tramite mappe di salienza (Selvaraju et al., 2017; Simonyan et al., 2013; Springenberg et al., 2014; Zeiler e Fergus, 2014; Zhou et al., 2016).

2.2 SPIEGAZIONE MODELLO-SPECIFICA E SPIEGAZIONE MODELLO-AGNOSTICA

La distinzione tra spiegazione modello-specifica e modello-agnostica è correlata a quella tra spiegazione basata sul modello e spiegazione post hoc (Adadi e Berrada, 2018), ma ci sono alcune differenze sfumate.

2.2.1 Spiegazione modello-specifica

I metodi di spiegazione specifici per il modello sono limitati a particolari classi di modelli. Ad esempio, un metodo di questo tipo può utilizzare attributi specifici per un tipo di rete neurale. Uno svantaggio è che, puntando a una spiegazione specifica del modello, si limita la scelta delle reti neurali, escludendo così potenzialmente una rete neurale che potrebbe adattare meglio l'output ai dati di input.

La spiegazione basata sul modello è per definizione specifica del modello (Adadi e Berrada, 2018), ma la spiegazione specifica del modello non è necessariamente basata sul modello. Alcune tecniche di mappatura della salienza post hoc sono esempi di tecniche specifiche per una certa classe di reti neurali convoluzionali (CNN), ma non sono metodi di spiegazione basati sul modello (Murdoch et al., 2019).

2.2.2 Spiegazione modello-agnostica

La spiegazione modello-agnostica è indipendente dalla scelta del tipo di rete neurale, operando esclusivamente sull'input e sull'output della rete neurale. Perturbando l'ingresso, l'utente può verificare quale sia il cambiamento nell'uscita della rete neurale. In questo modo è possibile spiegare quali regioni stanno

guidando l'uscita. Le spiegazioni "modello-agnostica" sono naturalmente post hoc.

2.3 AMBITO DI APPLICAZIONE DELLA SPIEGAZIONE

L'ambito di una spiegazione distingue tra la spiegazione di un intero modello (globale) e la spiegazione di un singolo output (locale).

2.3.1 Spiegazione globale

La spiegazione globale, detta anche spiegazione a livello di set di dati, fornisce relazioni generali apprese dalla rete neurale. Ad esempio, la spiegazione globale potrebbe fornire i punteggi di importanza delle caratteristiche a livello di set di dati, cioè quanto le caratteristiche contribuiscono all'output nell'intero set di dati (Olden et al., 2004). A titolo di esempio, si potrebbe osservare da una rete neurale che – o anche quanto – l'alta pressione sanguigna aumenta il rischio di un evento cardiaco. Un altro esempio di spiegazione globale potrebbe essere la visualizzazione dei filtri appresi, ossia quali caratteristiche vengono estratte dalla rete neurale e in che misura sono significative per il compito da svolgere (Olah et al., 2017; Zeiler e Fergus, 2014).

2.3.2 Spiegazione locale

La spiegazione locale fornisce una spiegazione di un singolo input. Nell'esempio del rischio cardiaco, l'input sarebbe una singola persona. La spiegazione locale spiegherebbe quindi perché la pressione arteriosa è importante per il rischio di eventi cardiaci per quella singola persona, mentre la spiegazione globale descriverebbe la relazione tra pressione arteriosa e rischio di eventi cardiaci nell'intero set di dati. Un altro esempio di spiegazione locale potrebbe essere una mappa di salienza che individua un tumore cerebrale su una risonanza magnetica (MRI) per spiegare quale parte della MRI ha contribuito principalmente al risultato del classificatore "tumore". Poiché spiega quale parte dell'imma-

gine ha portato il classificatore a produrre "tumore" per quella singola persona, si tratta di una spiegazione locale.

3. XAI NELL'ANALISI DELLE IMMAGINI MEDICHE

In questa sezione presenteremo quali tecniche XAI sono utilizzate nell'analisi delle immagini mediche e discuteremo gli adattamenti dei metodi tipicamente visti nella computer vision. Classifichiamo i metodi di spiegazione in tre tipi: visivi, testuali e basati su esempi; e classificheremo ogni metodo in base alla struttura di spiegazione basata sul modello rispetto a quella post-hoc, specifica rispetto a quella agnostica e globale rispetto a quella locale (Fig. 1). La Tabella 1 fornisce una panoramica delle tecniche più frequentemente utilizzate e mostra le loro connessioni secondo la tassonomia definita nella Sezione 2.

Tecnica	Sezione	Autori	Basato su modelli	Post hoc	Modello-specifico	Modello-agnostico	Globale	Locale
Spiegazione visiva	3.1.							
Approcci basati sulla retropropagazione	3.1.1							
Retropropagazione	3.1.1.1.	Simonyan et al. (2013)		✓	✓			✓
Deconvoluzione	3.1.1.1.	Zeiler e Fergus (2014)		✓	✓			✓
Retropropagazione guidata	3.1.1.1.	Springenberg et al. (2014)		✓	✓			✓
Mappatura dell'attivazione della classe (CAM)	3.1.1.2.	Zhou et al. (2016)		✓	✓			✓

Mappatura dell'attivazione della classe pesata con gradiente (Grad-CAM)	3.1.1.3.	Selvaraju et al. (2017)		✓	✓			✓
Propagazione della rilevanza a livello di strato (LRP)	3.1.1.4.	Bach et al. (2015)		✓	✓			✓
Esopiani additivi Deep SHapley (Deep SHAP)	3.1.1.5.	Lundberg e Lee (2017)		✓	✓	✓*	✓*	✓
Attenzione allenabile	3.1.1.6.	Jetley et al. (2018)	✓		✓			✓
Approcci basati sulla perturbazione	3.1.2							
Sensibilità all'occlusione	3.1.2.1.	Zeiler e Fergus (2014)		✓		✓		✓
Spiegazioni locali interpretabili e modello-agnostico (LIME)	3.1.2.2.	Ribeiro et al. (2016)		✓		✓		✓
Perturbazione significativa	3.1.2.3.	Fong e Vedaldi (2017)		✓		✓		✓
Analisi delle differenze di previsione	3.1.2.4.	Zintgraf et al. (2017)		✓		✓		✓
Spiegazione testuale	3.2.							
Didascalia dell'immagine	3.2.1.	Vinyals et al. (2015)	✓		✓			✓

Didascalia delle immagini con spiegazione visiva	3.2.2.	Zhang et al. (2017a)	✓		✓			✓
Test con vettori di attivazione del concetto (TCAV)	3.2.3.	Kim et al. (2018)		✓		✓	✓	✓
Spiegazione basata su esempi	3.3.							
Reti di triplette	3.3.1.	Hoffer e Ailon (2015)	✓		✓		✓	✓
Funzioni di influenza	3.3.2.	Wei Koh e Liang (2017)		✓		✓	✓	
Prototipi	3.3.3	Chen et al. 2019	✓		✓			✓

*Tabella 1. Panoramica delle tecniche di eXplainable AI (XAI) utilizzate nell'analisi delle immagini mediche, classificate in base al quadro di riferimento della Sezione 2. * Le ex-pianificazioni additive Shapley profonde sono post hoc e specifiche del modello a causa del metodo di ottimizzazione, ma le ex-pianificazioni additive Shapley possono anche essere globali e indipendenti dal modello.*

3.1 SPIEGAZIONE VISIVA

La spiegazione visiva, detta anche mappatura di salienza, è la forma più comune di XAI nell'analisi delle immagini mediche (Fig. 2). Le mappe di salienza mostrano le parti importanti di un'immagine per una decisione. La maggior parte delle tecniche di mappatura della salienza utilizza approcci basati sulla retropropagazione, ma alcune utilizzano approcci basati sulla perturbazione o sull'apprendimento di istanze multiple. Questi approcci saranno discussi di seguito. Nella Tabella 2 è riportata una panoramica dei lavori che utilizzano le mappe di salienza nell'imaging medico.

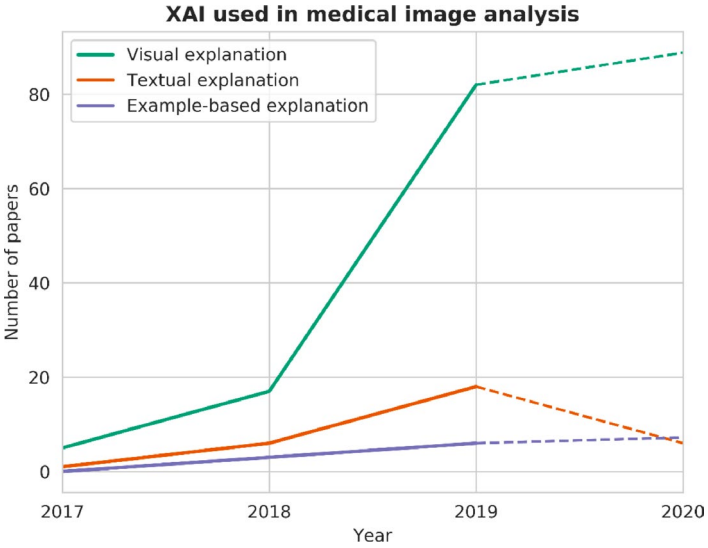


Fig. 2. Numero di articoli pubblicati ogni anno sull'analisi delle immagini mediche, per i tre tipi di tecniche XAI. La maggior parte degli articoli utilizza una spiegazione visiva. L'asse delle ordinate mostra il numero di articoli inclusi in questa indagine, mentre l'asse delle ascisse mostra l'anno in cui questi articoli sono stati pubblicati. La linea tratteggiata per il 2020 è un'estrapolazione data la situazione al 31 ottobre 2020.

Posizione anatomica	Autori (anno)	Modalità	Principale tecnica XAI utilizzata/ basata su
Vescica	Woerl et al. (2020)	Istologia	CAM
Cervello	Ahmad et al. (2019)	RISONANZA MAGNETICA	CAM
	Baumgartner et al. (2018)	RISONANZA MAGNETICA	CAM
	Böhle et al. (2019)	RISONANZA MAGNETICA	LRP
	Ceschin et al. (2018)	RISONANZA MAGNETICA	CAM
	Chakraborty et al. (2020)	RISONANZA MAGNETICA	CAM
	Choi et al. (2020)	PET/TC	CAM
	Dang e Chaudhury (2019)	RISONANZA MAGNETICA	LRP
	Dubost et al. (2019b)	RISONANZA MAGNETICA	Retropropagazione guidata
	Dubost et al. (2019a)	RISONANZA MAGNETICA	Sensibilità all'occlusione
	Dubost et al. (2020)	RISONANZA MAGNETICA	Attenzione allenabile
	Eitel et al. (2019)	RISONANZA MAGNETICA	LRP

	Fuchigami et al. (2020)	CT	Retropropagazione
	Gao et al. 2019	RISONANZA MAGNETICA	Deconvoluzione
	Gao et al. (2019)	RISONANZA MAGNETICA	CAM
	Grigorescu et al. (2019)	RISONANZA MAGNETICA	LRP
	Hilbert et al. (2019)	RISONANZA MAGNETICA	Grad-CAM
	Kim e Ye (2020)	RISONANZA MAGNETICA	Grad-CAM
	Kubach et al. (2020)	Istologia	Grad-CAM guidata
	Lee et al. (2019b)	CT	CAM
	Li et al. 2019b	RISONANZA MAGNETICA	CAM
	Lian et al. (2019)	RISONANZA MAGNETICA	Attenzione allenabile
	Liao et al. (2020)	RISONANZA MAGNETICA	Grad-CAM
	Lin et al. (2019)	Ultrasuoni	CAM
	Natekar et al. (2020)	RISONANZA MAGNETICA	Grad-CAM
	Ng et al. (2018)	RISONANZA MAGNETICA	CAM
	Pereira et al. (2018)	RISONANZA MAGNETICA	Grad-CAM
	Pominova et al. (2018)	RISONANZA MAGNETICA	Grad-CAM
	Rezaei et al. (2020)	RISONANZA MAGNETICA	Retropropagazione
	Saab et al. (2019)	CT	Apprendimento a istanze multiple
	Seo et al. (2020)	RISONANZA MAGNETICA	Analisi delle differenze di previsione
	Shahamat e Saniee Abadeh (2020)	RISONANZA MAGNETICA	Sensibilità all'occlusione
	Shinde et al. (2019a)	RISONANZA MAGNETICA	CAM
	Shinde et al. (2019b)	RISONANZA MAGNETICA	CAM
	Tang et al. 2019	Istologia	Grad-CAM
	Wang et al. 2020c	RISONANZA MAGNETICA	Retropropagazione guidata
	Wei et al. (2019)	RISONANZA MAGNETICA	Retropropagazione
	Windisch et al. (2020)	RISONANZA MAGNETICA	Grad-CAM
	Xie et al. (2020)	Ultrasuoni	Grad-CAM
	Xu et al. (2019)	RISONANZA MAGNETICA	Attenzione allenabile
	Xu et al. (2019)	RISONANZA MAGNETICA	LRP
	Ye et al. (2019)	CT	Grad-CAM
	Zintgraf et al. (2017)	RISONANZA MAGNETICA	Analisi delle differenze di previsione
Seno	Akselrod-Ballin et al. (2019)	Raggi X	Perturbazione significativa
	El Adoui et al. (2020)	RISONANZA MAGNETICA	Grad-CAM
	Gecer et al. (2018)	Istologia	Sensibilità all'occlusione

	Huang et al. (2020)	Raggi X	CAM
	Kim et al. (2020)	Ultrasuoni	CAM
	Lee e Nishikawa (2019)	Raggi X	CAM
	Luo et al. (2019)	RISONANZA MAGNETICA	CAM
	Maicas et al. (2019)	RISONANZA MAGNETICA	Apprendimento a istanze multiple
	Obikane e Aoki (2020)	Istologia	Grad-CAM
	Papanastasopoulos et al. (2020)	RISONANZA MAGNETICA	Gradiente integrato
	Qi et al. (2019)	Ultrasuoni	CAM
	van der Velden et al. (2020)	RISONANZA MAGNETICA	SHAP
	Wang et al. (2018)	Raggi X	Attenzione allenabile
	Xi et al. (2019)	Raggi X	CAM
	Yang et al. (2019)	Istologia	Attenzione allenabile
	Yi et al. (2019)	Raggi X	CAM
	Zhou et al. 2020	Ultrasuoni	CAM
Cardiovascolare	Candemir et al. (2020)	CT	Grad-CAM
	Cong et al. (2019)	Raggi X	Grad-CAM
	Gessert et al. (2019)	OCT	Retropropagazione guidata
	Huo et al. (2019)	CT	Grad-CAM
	Patra e Noble (2020)	Ultrasuoni	Grad-CAM
	de Vos et al. (2019)	CT	Deconvoluzione
Petto	Ausawalaithong et al. (2018)	Raggi X	CAM
	Brunese et al. (2020)	Raggi X	Grad-CAM
	Chen et al. (2019)	Raggi X	Grad-CAM
	Dunnmon et al. (2019)	Raggi X	CAM
	Guo et al. (2020)	CT	CAM
	He et al. (2017)	Istologia	Grad-CAM
	Hosny et al. (2018)	CT	Grad-CAM
	Huang e Fu (2019)	Raggi X	CAM
	Humphries et al. (2020)	CT	Grad-CAM
	Khakzar et al. (2019)	Raggi X	CAM
	Ko et al. (2020)	CT	Grad-CAM
	Kumar et al. (2019a)	CT	CAM
	Lei et al. (2020)	CT	CAM
	Li et al. 2019d	Raggi X	Apprendimento a istanze multiple
	Liu et al. 2019f	Raggi X	CAM
	Mahmud et al. (2020)	Raggi X	Grad-CAM

	Paul et al. 2020	CT	Grad-CAM
	Pesce et al. (2019)	Raggi X	Attenzione allenabile
	Philbrick et al. (2018)	CT	Grad-CAM
	Qin et al. (2020)	PET/TC	Grad-CAM
	Rajaraman et al. (2019)	Raggi X	LIME
	Rajpurkar et al. (2018)	Raggi X	CAM
	Schwab et al. (2020)	Raggi X	Apprendimento a istanze multiple
	Sedai et al. (2018)	Raggi X	CAM
	Singla et al. (2018)	CT	Attenzione allenabile
	Tang et al. (2019)	CT	CAM
	Tang et al. (2020)	Raggi X	CAM
	Teramoto et al. (2019)	Istologia	Grad-CAM
	van Sloun e Demi (2019)	Ultrasuoni	Grad-CAM
	Wang et al. 2019	Raggi X	CAM
	Xu et al. 2019	CT	Grad-CAM
	Paul et al. (2020)	Raggi X	CAM
	Zhu e Ogino (2019)	CT	SHAP
Odontoiatrico	Vila-Blanco et al. (2020)	Raggi X	Grad-CAM
Occhio	Ahmad et al. 2019	Fotografia del fundus	CAM
	Araújo et al. (2020)	Fotografia del fundus	Apprendimento a istanze multiple
	Costa et al. (2019)	Fotografia del fundus	Apprendimento a istanze multiple
	Jang et al. (2018)	Fotografia del fundus	Grad-CAM guidata
	Jiang et al. (2019)	Fotografia del fundus	CAM
	Kim et al. (2019)	Fotografia del fundus	Grad-CAM
	Kumar et al (2019b)	Fotografia del fundus	CAM
	Li et al. 2019a	Fotografia del fundus	Attenzione allenabile
	Liao et al. (2019)	Fotografia del fundus	CAM
	Liu et al. (2019)	Fotografia del fundus	CAM
	Martins et al. (2020)	Fotografia del fundus	Grad-CAM
	Meng et al. (2020)	Fotografia del fundus	Grad-CAM
	Narayanan et al. (2020)	Fotografia del fundus	CAM
	Perdomo et al. (2019)	OCT	CAM
	Quellec et al. (2020)	Fotografia del fundus	Retropropagazione
	Shen et al. (2020)	Fotografia del fundus	CAM
	Thakoor et al. (2019)	OCT	Grad-CAM
	Tu et al. (2020)	Fotografia del fundus	CAM

	Wang et al. 2020a	OCT	Grad-CAM
	Wang et al. 2020b	CT	CAM
	Wang et al. 2019b	Fotografia del fundus	CAM
	Zhang et al. (2019)	Fotografia del fundus	Grad-CAM
	Zhou et al. (2020)	OCT	CAM
Sistema riproduttivo femminile	Gupta et al. (2020)	Istologia	Grad-CAM
	GV e Reddy (2019)	Istologia	Grad-CAM
	Sun et al. (2020)	Istologia	CAM
Gastrointestinale	Chen et al. 2019	CT	Grad-CAM
	Everson et al. (2019)	Endoscopia	CAM
	García-Peraza-Herrera et al. (2020)	Endoscopia	CAM
	Heinemann et al. (2019)	Istologia	CAM
	Itoh et al. (2020)	Endoscopia	Grad-CAM
	Kiani et al. (2020)	Istologia	CAM
	Korbar et al. (2017)	Istologia	Grad-CAM
	Kowsari et al. (2020)	Istologia	Grad-CAM
	Lee et al. 2020	Ultrasuoni	Retropropagazione
	Malhi et al. (2019)	Endoscopia	LIME
	Rajpurkar et al. (2020b)	CT	Grad-CAM
	Shapira et al. (2020)	CT	Apprendimento a istanze multiple
	Wang et al. (2020)	RISONANZA MAGNETICA	Grad-CAM
	Wang et al. 2019a	Endoscopia	CAM
	Wickstrøm et al. (2020)	Endoscopia	Retropropagazione guidata
	Yan et al. (2020)	Istologia	CAM
	Zhu et al. (2020)	Istologia	Attenzione allenabile
Linfonodi	Ji (2019)	Istologia	Grad-CAM
Muscoloscheletrico	Bien et al. (2018)	RISONANZA MAGNETICA	CAM
	Chang et al. (2020)	RISONANZA MAGNETICA	CAM
	Cheng et al. (2019)	Raggi X	Grad-CAM
	Gupta et al. 2020	Raggi X	Grad-CAM
	Jamaludin et al. (2017)	RISONANZA MAGNETICA	Retropropagazione guidata
	Kim et al. 2020	Raggi X	Retropropagazione
	Paul et al. (2019)	Raggi X	CAM
	Zhang et al. (2020)	Raggi X	Grad-CAM
	Zhao et al. (2018)	Raggi X	CAM

	von Schacky et al. (2020)	Raggi X	Grad-CAM
Prostata	Silva-Rodríguez et al. (2020)	Istologia	CAM
	Yang et al. (2017)	RISONANZA MAGNETICA	CAM
La pelle	Barata et al. (2020)	Dermatoscopia	Attenzione allenabile
	Bian et al. (2019)	Fotografia	Retropropagazione
	Li et al. (2020)	Dermatoscopia	CAM
	Li et al. 2019c	Fotografia	Analisi delle differenze di previsione
	Xie et al. 2020	Fotografia	CAM
	Yan et al. 2019	Dermatoscopia	Attenzione allenabile
	Young et al. (2019)	Dermatoscopia	SHAP
	Zunair e Hamza (2020)	Fotografia	Grad-CAM
Cranio	Kim et al. 2019b	Raggi X	CAM
Tiroide	Lee et al. (2020)	CT	Grad-CAM
	Wang et al. 2019	Ultrasuoni	Attenzione
	Wang et al. 2020	Ultrasuoni	CAM
Multiplo	Chan et al. (2019)	Istologia	Grad-CAM
	Huang e Chung (2019)	Istologia	CAM
	Hägele et al. (2020)	Istologia	LRP
	Kermany et al. (2018)	Multiplo	Sensibilità all'occlusione
	Kim et al. 2019	Multiplo	CAM
	Langner et al. (2019)	RISONANZA MAGNETICA	Grad-CAM
	Meng et al. (2019)	Ultrasuoni	Attenzione allenabile
	Schlemper et al. (2019)	CT	Attenzione allenabile
	Tang (2020)	Multiplo	CAM
	Upadhyay e Banerjee (2020)	Multiplo	Grad-CAM

Tabella 2. Papers che forniscono spiegazioni visive. Per motivi di leggibilità, i papers sono ordinati in base alla posizione anatomica e solo il primo documento che tratta quella posizione anatomica riporta il nome della posizione. La colonna "Principale tecnica XAI utilizzata/basata su" descrive quale tecnica di spiegazione visiva della Sezione 3.1 è stata utilizzata o su quale tecnica si basa il metodo del documento corrispondente. Quando sono state applicate più tecniche di spiegazione visiva, è stata annotata la tecnica più recente in base alla Tabella 1. CAM = class activation mapping [mappatura dell'attivazione della classe], CT = computed tomography [tomografia computerizzata], LIME = local interpretable model-agnostic explanations [spiegazioni locali interpretabili e agnostiche del modello], LRP = Layer-wise relevance propagation [Propagazione della rilevanza a livello di strato], MRI = magnetic resonance imaging [risonanza magnetica] , OCT = optical coherence tomography [tomografia a coerenza ottica], PET = positron emission tomography [tomografia a emissione di positroni], SHAP = Shapley additive explanations [Spiegazioni additive di Shapley].

3.1.1 Approcci basati sulla retropropagazione

(Guided) backpropagation *e* deconvolution** *[Retropropagazione (guidata) e deconvoluzione]*: Alcune delle prime tecniche per creare mappe di salienza hanno evidenziato i pixel che avevano il maggiore impatto sull'output dell'analisi. Tra gli esempi vi sono la visualizzazione delle derivate parziali dell'output a livello di pixel (Simonyan et al., 2013), la deconvoluzione (Zeiler e Fergus, 2014) e la retropagazione guidata (Springenberg et al., 2014). Queste tecniche forniscono spiegazioni post hoc locali, specifiche per il modello (solo per le CNN). Queste tecniche sono state utilizzate nell'analisi delle immagini mediche. Ad esempio, de Vos et al. (2019) hanno stimato la quantità di calcio coronarico per ogni fetta di immagine di tomografia computerizzata (TC) cardiaca o toracica e hanno utilizzato la deconvoluzione per visualizzare da quale punto della fetta si è basata la decisione.

Mappature dell'attivazione della classe (CAM): Zhou et al. (2016) hanno introdotto il Class Activation Mapping (CAM). Hanno sostituito gli strati completamente connessi alla fine di una CNN con un pooling della media globale sulle ultime mappe di caratteristiche convoluzionali. La mappa di attivazione della classe era una somma lineare ponderata della presenza di modelli visivi (catturati dai filtri) in diverse posizioni spaziali. Questa tecnica fornisce una spiegazione locale, specifica per il modello e post hoc. Diversi ricercatori hanno utilizzato questa tecnica nell'imaging medico (Tabella 2).

Le CAM sono state utilizzate anche nell'analisi di immagini mediche in ensemble di CNN. Ad esempio, Jiang et al. (2019) hanno costruito un ensemble di Inception-V3, ResNet-152 e Inception-ResNet-V2 per distinguere le immagini del fundus di soggetti sani o di pazienti con retinopatia diabetica lieve da quelle con retinopatia diabetica moderata o grave e hanno fornito una combinazione ponderata delle CAM risultanti per la localizzazione della retinopatia diabetica. Lee et al. (2019b) hanno costruito CAM dall'output di un insieme di quattro CNN:

VGG-16, ResNet-50, Inception-V3 e Inception-ResNet-V2, per il rilevamento di emorragie intracraniche acute.

Poiché le immagini mediche spesso contengono informazioni a più scale, sono state proposte anche CAM multiscala. Liao et al. (2019) hanno concatenato mappe di caratteristiche a tre scale che sono state successivamente fornite come input per il pooling della media globale. Le mappe di attivazione ottenute hanno mostrato una risoluzione più elevata rispetto alle mappe a scala singola e sono risultate migliori nell'identificazione di piccole strutture sulle immagini del fundus della retina. Shinde et al. (2019a) hanno concatenato le mappe delle caratteristiche di ogni strato prima del max-pooling e le hanno fornite come input a uno strato di pooling medio globale. Le loro CAM ad "alta risoluzione" hanno fornito localizzazioni accurate di tumori cerebrali alla risonanza magnetica. García-Peraza-Herrera et al. (2020) hanno proposto l'estrazione di CAM a più risoluzioni. Hanno dimostrato che i CAM ad alta risoluzione erano accurati nell'evidenziare i modelli di anse capillari interpapillari nelle immagini endoscopiche, che erano relativamente piccole rispetto all'intera immagine.

Mappatura dell'attivazione della classe ponderata per gradi (Grad-CAM): Selvaraju et al. (2017) hanno introdotto il Gradient-weighted Class Activation Mapping (Grad-CAM), che è una generalizzazione del CAM. Grad-CAM può funzionare con qualsiasi tipo di CNN per produrre spiegazioni locali post hoc, mentre CAM necessita specificamente di un pooling della media globale. Gli autori hanno anche introdotto la Grad-CAM guidata, una moltiplicazione elementare tra la retropagazione guidata e la Grad-CAM. Grad-CAM e Guided Grad-CAM sono stati utilizzati nell'analisi delle immagini mediche. Ad esempio, Ji (2019) ha utilizzato Grad-CAM per mostrare su quali aree di sezioni linfonodali istologiche un classificatore ha basato la sua decisione di tessuto metastatico; Kowsari et al. (2020) lo hanno utilizzato per individuare le enteropatie del piccolo intestino sull'istologia; e Windisch et al. (2020) hanno utilizzato Grad-Cam per mostrare

quali aree della risonanza magnetica cerebrale hanno fatto sì che il classificatore decidesse sulla presenza di un tumore.

Propagazione della rilevanza a livello di strato (LRP): Bach et al. (2015) hanno introdotto la layer-wise relevance propagation (LRP). LRP utilizza l'output della rete neurale, ad esempio un punteggio di classificazione compreso tra 0 e 1, e lo retropropaga iterativamente in tutta la rete. In ogni iterazione (cioè in ogni strato), LRP assegna un punteggio di rilevanza a ciascuno dei neuroni di ingresso degli strati precedenti. Questi punteggi di rilevanza distribuiti devono essere uguali al punteggio di rilevanza totale del neurone di origine, secondo la legge di conservazione.

LRP è stato utilizzato nell'analisi delle immagini mediche. Ad esempio, Böhle et al. (2019) hanno utilizzato LRP per identificare le regioni responsabili della malattia di Alzheimer dalle immagini di risonanza magnetica del cervello. Hanno confrontato le mappe di salienza fornite da LRP con quelle fornite dalla retropagazione guidata e hanno scoperto che LRP era più specifico nell'identificare le regioni note per la malattia di Alzheimer.

Deep SHapley Additive exPlanations (Deep SHAP): Lundberg e Lee (2017) hanno proposto un approccio unificato per spiegare le previsioni utilizzando SHapley Additive exPlanations (SHAP). Questo approccio indipendente dal modello utilizza i valori di Shapley (Shapley, 2016), un concetto tratto dalla teoria dei giochi. I valori di Shapley determinano il contributo marginale di ogni caratteristica all'output del modello individualmente. Un aspetto negativo dei valori di Shapley è che il loro calcolo richiede molte risorse, poiché richiede la valutazione di molte permutazioni.

Combinando DeepLIFT con i valori Shapley, Lundberg e Lee (2017) hanno proposto un metodo veloce per approssimare i valori Shapley per le CNN, chiamato Deep SHAP. Deep SHAP è stato utilizzato nell'analisi delle immagini mediche. Ad esempio, van der Velden et al. (2020) hanno utilizzato una CNN di regressione per stimare la densità volumetrica del seno dalla risonanza magnetica mammaria. Deep SHAP è stato utilizzato per spiegare

quali parti dell'immagine hanno un contributo positivo e quali negativo alla stima della densità.

Attenzione addestrabile: Mentre molte delle tecniche menzionate in precedenza hanno evidenziato quali sono le regioni dell'immagine su cui si concentra la rete, cioè dove è diretta l'attenzione, Jetley et al. (2018) hanno proposto un meccanismo di attenzione addestrabile. Questo metodo di attenzione addestrabile ha evidenziato dove e in quale proporzione la rete ha prestato attenzione alle immagini di input per la classificazione e ha utilizzato questa attenzione per amplificare ulteriormente le aree rilevanti e sopprimere quelle irrilevanti.

Nell'imaging medico, Schlemper et al. (2019) hanno utilizzato l'attenzione addestrabile e introdotto l'attenzione a griglia. La logica alla base di questa scelta è che la maggior parte degli oggetti di interesse nelle immagini mediche sono altamente localizzati. Utilizzando l'attenzione a griglia, l'attenzione addestrabile ha catturato le informazioni anatomiche nelle immagini mediche. Hanno dimostrato prestazioni elevate sia per la segmentazione che per la localizzazione, aggiungendo le porte di attenzione a UNET (Ronneberger et al., 2015) e a una variante di VGG (Simonyan e Zisserman, 2014). I coefficienti di attenzione sono stati utilizzati per spiegare su quali aree dell'immagine si è concentrata la rete.

3.1.2 Approcci basati sulla perturbazione

3.1.2.1 Sensibilità all'occlusione

Le tecniche basate sulla perturbazione perturbano l'immagine di input per valutare l'importanza di determinate aree dell'immagine per il compito in esame. Zeiler e Fergus (2014) hanno utilizzato un'analisi di sensibilità all'occlusione per visualizzare quali parti dell'immagine fossero più importanti per la classificazione. Ad esempio, hanno dimostrato che l'immagine di un cane che tiene in mano una pallina da tennis veniva classificata correttamente in base alla razza del cane, tranne quando il volto

del cane era occluso, il che produceva la classificazione errata "pallina da tennis".

3.1.2.2 Spiegazioni locali interpretabili modello-agnostiche (LIME)

Ribeiro et al. (2016) hanno introdotto le Local Interpretable Model-agnostic Explanations [Spiegazioni locali interpretabili e agnostiche del modello] (LIME). Le LIME forniscono spiegazioni locali sostituendo localmente un modello complesso con modelli più semplici, ad esempio approssimando una CNN con un modello lineare. Perturbando i dati di ingresso, l'output del modello complesso cambia. LIME utilizza il modello più semplice per apprendere la mappatura tra i dati di ingresso perturbati e la variazione dell'output. La somiglianza tra l'input perturbato e l'input originale viene utilizzata come peso, per garantire che le spiegazioni fornite dai modelli semplici con input altamente perturbati abbiano un effetto minore sulla spiegazione finale. Nelle immagini, Ribeiro et al. (2016) hanno implementato le perturbazioni utilizzando superpixelxxxxxxxxx è incluso nel collegamento ipertestuale"?> Achanta et al., 2012), piuttosto che singoli pixel, per mostrare quali regioni erano importanti per spiegare una classificazione.

Il LIME è stato utilizzato da diversi ricercatori nell'analisi delle immagini mediche. Ad esempio, Malhi et al. (2019) hanno utilizzato il LIME per spiegare quali aree nelle immagini di endoscopia gastrica contenevano regioni sanguinolente.

3.1.2.3 Perturbazione significativa

Fong e Vedaldi (2017) hanno introdotto la perturbazione significativa, in cui hanno perturbato l'immagine di ingresso per rilevare i cambiamenti nelle previsioni di una rete neurale addestrata. Piuttosto che utilizzare perturbazioni come la sensibilità all'occlusione che blocca parti dell'immagine, hanno suggerito di simulare effetti naturalistici o plausibili, che portano a perturbazioni più significative e di conseguenza a spiegazioni più significative. Hanno optato per tre tipi di perturbazioni locali, ossia un valore costante, un rumore o una sfocatura.

Uzunova et al. (2019) hanno affermato che le perturbazioni proposte da Fong e Vedaldi (2017) non sono adatte alle immagini mediche. Sostituire aree di un'immagine medica con un valore costante non è plausibile e le immagini mediche tendono naturalmente a essere rumorose e sfocate. Hanno proposto di sostituire le regioni patologiche con un equivalente di tessuto sano utilizzando un autoencoder variazionale (VAE). Hanno dimostrato che le perturbazioni del VAE individuano le regioni patologiche in diversi studi di imaging, come le immagini di tomografia a coerenza ottica dell'occhio (la patologia consisteva in fluido intraretinico, fluido sottoretinico e distacchi dell'epitelio pigmentato) e la risonanza magnetica del cervello (la patologia consisteva in lesioni da ictus). Inoltre, hanno dimostrato che l'uso di una VAE ha prodotto una migliore localizzazione della patologia rispetto all'uso di una semplice sfocatura o di perturbazioni a valore costante.

Lenis et al. (2020) hanno utilizzato un ragionamento simile a quello di Uzunova et al. (2019) e hanno usato l'inpainting per sostituire le regioni patologiche con equivalenti di tessuto sano. Hanno dimostrato che le perturbazioni create dall'inpainting superano le prestazioni di backpropagation e Grad-CAM nell'individuazione di masse nella mammografia e di tubercolosi nelle radiografie del torace, sulla base della distanza di Hausdorff tra le heatmap sogliate derivate dalle mappe di salienza e le etichette di valori di base a livello di pixel.

3.1.2.4 Analisi delle differenze di previsione

Zintgraf et al. (2017) hanno adattato l'analisi della differenza di predizione (Robnik-Šikonja e Kononenko, 2008) per generare mappe di salienza. Se ogni pixel di un'immagine è considerato una caratteristica, l'analisi della differenza di predizione assegna un valore di rilevanza a ogni pixel, misurando come cambia la predizione se il pixel è considerato sconosciuto. Zintgraf et al. (2017) hanno ampliato questo approccio aggiungendo il campionamento condizionale, ovvero analizzando solo i pixel difficili da prevedere semplicemente analizzando i pixel vicini, e aggiun-

gendo l'analisi multivariabile, ovvero analizzando patch di pixel connessi invece di singoli pixel. Hanno incluso un'analisi della risonanza magnetica cerebrale dei pazienti affetti da HIV rispetto ai controlli sani, ottenendo una spiegazione della decisione del classificatore.

Seo et al. (2020) hanno utilizzato l'analisi delle differenze di predizione in combinazione con superpixel (o supervoxel per il 3D) su più scale. Queste mappe di salienza basate su supervoxel multiscala hanno fornito spiegazioni che gli autori hanno descritto come visivamente piacevoli, poiché seguono i bordi dell'immagine. Le mappe di salienza spiegavano quali regioni erano informative per un classificatore per distinguere tra pazienti con malattia di Alzheimer e controlli normali.

3.1.3 Approcci basati sull'apprendimento di più istanze

L'apprendimento a istanze multiple può essere utilizzato per visualizzare le spiegazioni. Nell'apprendimento a istanze multiple, gli insiemi di allenamento sono costituiti da sacche di istanze (Dietterich et al., 1997). Queste sacche sono etichettate, ma le istanze non lo sono. Nell'analisi delle immagini mediche, l'apprendimento di istanze multiple può essere effettuato, ad esempio, utilizzando un approccio basato sulle patch: Un'immagine rappresenta la sacca e le patch di quell'immagine rappresentano le istanze (Cheplygina et al., 2019).

Diversi ricercatori hanno utilizzato questo approccio per individuare le istanze del sacchetto responsabili della classificazione. Ad esempio, Schwab et al. (2020) hanno localizzato i reperti critici nelle radiografie del torace utilizzando un approccio basato sulle patch. Ogni patch dell'immagine ha ricevuto una previsione e le previsioni sono state sovrapposte all'immagine per visualizzare su quali aree il classificatore ha basato la sua decisione. Araújo et al. (2020) hanno utilizzato l'apprendimento a istanze multiple per spiegare quali aree di una fotografia del fundus sono importanti per la retinopatia diabetica. Hanno valutato la gravità della malattia usando una scala ordinale con gradi da 0 a

5. Utilizzando un approccio basato su patch, hanno fornito mappe di spiegazione visiva per ogni grado di retinopatia diabetica.

3.2 SPIEGAZIONE TESTUALE

La spiegazione testuale è una forma di XAI che fornisce descrizioni testuali. Tali descrizioni includono caratteristiche relativamente semplici (ad esempio, "massa spicata"), fino a interi referti medici. Descriveremo tre tipi di spiegazione testuale: didascalia delle immagini, didascalia delle immagini con spiegazione visiva e test con attribuzione di concetti.

La tabella 3 mostra una panoramica dei lavori che utilizzano la spiegazione testuale nell'imaging medico.

Posizione anatomica	Autori (anno)	Modalità	Principale tecnica XAI utilizzata/basata su
Vescica	Zhang et al. (2017b)	Istologia	Didascalia delle immagini con spiegazione visiva
Seno	Kim et al. 2019a	Raggi X	Didascalia delle immagini con spiegazione visiva
	Lee et al. (2019a)	Raggi X	Didascalia delle immagini con spiegazione visiva
	Sun et al. (2019)	Raggi X	Didascalia dell'immagine
Cardiovascolare	Clough et al. (2019)	RISONANZA MAGNETICA	TCAV
Petto	Gasimova (2019)	Raggi X	Didascalia dell'immagine
	Kashyap et al. (2020)	Raggi X	Didascalia delle immagini con spiegazione visiva
	Li et al. (2019)	Raggi X	Didascalia delle immagini con spiegazione visiva
	Nunes et al. (2019)	Raggi X	Didascalia delle immagini con spiegazione visiva
	Rodin et al. (2019)	Raggi X	Didascalia delle immagini con spiegazione visiva
	Shen et al. (2019)	CT	Altre spiegazioni testuali
	Singh et al. (2019)	Raggi X	Didascalia dell'immagine
	Spinks e Moens (2019)	Raggi X	Didascalia dell'immagine
	Tian et al. (2019)	Raggi X	Didascalia dell'immagine

	Wang et al. 2019c	Raggi X	Didascalia delle immagini con spiegazione visiva
	Wu et al. (2018)	CT	TCAV
	Yan et al. (2019)	CT	Altre spiegazioni testuali
	Yang et al. 2020	Raggi X	Didascalia dell'immagine
	Yin et al. (2019)	Raggi X	Didascalia dell'immagine
	Yuan et al. (2019)	Raggi X	Didascalia delle immagini con spiegazione visiva
Occhio	Kim et al. (2018)	Fotografia del fundus	TCAV
Sistema riproduttivo femminile	Ma et al. (2018)	Istologia	Didascalia delle immagini con spiegazione visiva
Gastrointestinale	Tian et al. (2018)	CT	Didascalia delle immagini con spiegazione visiva
Rene	Maksoud et al. (2019)	Istologia	Didascalia dell'immagine
Muscoloscheletrico	Koitka et al. (2020)	Raggi X	Didascalia dell'immagine
Multiplo	Allaouzi et al. (2018)	Multiplo	Didascalia dell'immagine
	Graziani et al. (2020)	Multiplo	TCAV
	Jing et al. 2018	Multiplo	Didascalia delle immagini con spiegazione visiva
	Pelka et al. (2019)	Raggi X	Didascalia dell'immagine
	Zeng et al. (2020)	Multiplo	Didascalia dell'immagine

Tabella 3. Papers che forniscono spiegazioni testuali. Per motivi di leggibilità, i papers sono ordinati in base alla posizione anatomica e solo il primo paper che tratta quella posizione anatomica riporta il nome della posizione. La colonna "Principale tecnica XAI utilizzata/basata su" descrive quale tecnica di spiegazione testuale della Sezione 3.2 è stata utilizzata o su quale tecnica si basa il metodo del paper corrispondente. CT = tomografia computerizzata, TCAV = test con vettori di attivazione concettuale

3.2.1 Didascalia dell'immagine

Vinyals et al. (2015) hanno fornito spiegazioni testuali per le immagini utilizzando un framework di image captioning end-to-end. Hanno accoppiato una rete neurale convoluzionale per la codifica dell'immagine e una rete neurale ricorrente, in particolare una rete a memoria a breve termine (LSTM) (Hochreiter e Schmidhuber, 1997), per la codifica testuale. Per l'addestramento hanno utilizzato frasi generate dall'uomo come verità di base

e per la valutazione hanno utilizzato la metrica della valutazione bilingue (BLEU). La metrica BLEU descrive la precisione degli N-grammi di parole, cioè una sequenza di N parole, tra le frasi generate e quelle di riferimento (Papineni et al., 2002).

Singh et al. (2019) hanno utilizzato un framework per la didascalia delle immagini per fornire spiegazioni testuali sulle radiografie del torace. Hanno utilizzato i database di word-embedding Global Vectors (GloVe) (Pennington et al., 2014) e la variante radiologica RadGloVe (Zhang et al., 2018) per addestrare l'LSTM e hanno usato la metrica BLEU già citata e le varianti METEOR, CIDER e ROUGE (Banerjee e Lavie, 2005; Lin, 2004; Vedantam et al., 2015). Come previsto, sono state raggiunte prestazioni più elevate nel referto radiologico generato quando sono stati utilizzati sia RadGloVe che GloVe invece del solo GloVe.

3.2.2 Didascalia delle immagini con spiegazione visiva

Diversi ricercatori hanno combinato la didascalia delle immagini con la spiegazione visiva. Zhang et al. (2017a) hanno introdotto un framework che utilizza la doppia attenzione, sia per il testo che per le immagini. Hanno utilizzato un approccio simile a quello della didascalia delle immagini, ossia un codificatore per l'immagine e un LSTM per il testo, ma hanno aggiunto la doppia attenzione. Questo ha facilitato le interazioni di alto livello tra le previsioni dell'immagine e del testo e ha prodotto mappe di attenzione visiva corrispondenti alle spiegazioni testuali nelle immagini istologiche.

Wang et al. 2018 hanno utilizzato un approccio simile e hanno dimostrato, nel loro esempio di radiografia del torace, che diverse parti della spiegazione testuale portavano a diverse aree di mappatura della salienza nell'immagine. Hanno mostrato una mappa di salienza del torace con più regioni corrispondenti a diversi risultati radiologici.

Lee et al. (2019a) hanno mostrato la didascalia di immagini con spiegazione visiva per le mammografie. Hanno aggiunto una perdita di vincoli visivi alle LSTM che generano testo, per garan-

tire che le spiegazioni fornite seguano il gergo corretto dei referti mammografici. Hanno dimostrato che l'aggiunta di questa perdita aiuta a generare spiegazioni testuali migliori. Inoltre, hanno collegato i referti radiologici a mappe di salienza visiva.

3.2.3 Test con vettori di attivazione del concetto (TCAV)

Le attribuzioni concettuali forniscono spiegazioni corrispondenti a concetti di alto livello che gli esseri umani trovano facili da comprendere (Kim et al., 2018). Utilizzando il Testing with Concept Activation Vectors (TCAV), Kim et al. (2018) hanno presentato spiegazioni lineari a misura d'uomo dello stato interno delle reti neurali, ottenendo una spiegazione globale delle reti in termini di concetti comprensibili all'uomo. Questi concetti possono essere forniti dopo l'addestramento della rete neurale come analisi post hoc. L'algoritmo TCAV utilizza insiemi definiti dall'utente di esempi di un concetto e di esempi casuali di non concetti. Un concetto di questo tipo potrebbe essere "strisce" per valutare se un'immagine contiene una zebra, o "massa spicata" per valutare se un'immagine contiene un cancro. TCAV quantifica la sensibilità di un modello addestrato a tali concetti utilizzando i vettori di attivazione dei concetti (CAV). La risposta dei casi di test a questi CAV è stata poi utilizzata per misurare la sensibilità a quel concetto. Gli autori hanno dimostrato la fattibilità di TCAV su un esempio di elaborazione di immagini mediche, mettendo in relazione annotazioni mediche come "microaneurisma" con la retinopatia diabetica nell'imaging del fundus.

Clough et al. (2019) hanno identificato la patologia cardiaca nella risonanza magnetica per immagini classificando lo spazio latente di una VAE. Hanno utilizzato la TCAV per mostrare quali biomarcatori clinicamente noti erano correlati alla malattia cardiaca. Inoltre, hanno ricostruito le immagini con una bassa velocità di eiezione di picco – una caratteristica che potrebbe essere correlata a una malattia cardiaca – aggiungendo la CAV allo spazio latente.

Graziani et al. (2020) hanno ampliato il TCAV introducendo i vettori concettuali di regressione. L'aggiunta principale era che, mentre la TCAV indicava la presenza o l'assenza di concetti binari, i vettori concettuali di regressione indicavano misure a valore continuo di un concetto. Questo può essere utile quando si indaga su un concetto continuo come la dimensione del tumore. Graziani et al. (2020) hanno dimostrato che, utilizzando i vettori concettuali di regressione, potevano ad esempio spiegare perché la rete classificava un'area di un'immagine istopatologica del seno come tumorale e un'altra come sana: Entrambe le aree dell'immagine hanno ottenuto un punteggio elevato per il concetto di "contrasto", ma il concetto di "area dei nuclei", che si riferisce a un sistema clinicamente utilizzato per valutare le dimensioni delle cellule, era diverso tra le regioni sane e quelle cancerose.

3.2.4 Altre tecniche di spiegazione "tel"

Shen et al. (2019) hanno utilizzato quella che hanno definito una CNN semantica gerarchica per prevedere la malignità dei noduli polmonari alla TC. Hanno classificato cinque descrizioni testuali di caratteristiche dell'immagine rappresentative della malignità dei noduli polmonari, tipicamente valutate da un radiologo. Il compito di trovare le descrizioni testuali è stato combinato con il compito principale di classificare la malignità dei noduli polmonari. Sebbene la CNN semantica gerarchica non abbia superato in modo significativo una normale CNN nel predire la malignità dei noduli, il metodo ha fornito caratteristiche dei noduli interpretabili dall'uomo.

3.3 SPIEGAZIONE BASATA SU ESEMPI

La spiegazione basata su esempi è una tecnica XAI che fornisce esempi relativi al punto di dati che si sta analizzando. Questa tecnica può essere utile quando si cerca di spiegare perché una rete neurale è giunta a una decisione ed è correlata al modo in cui gli esseri umani ragionano. Ad esempio, quando un patologo

esamina una biopsia di un paziente che mostra una somiglianza con un paziente esaminato in precedenza dallo stesso patologo, la decisione clinica può essere migliorata conoscendo la valutazione della biopsia precedente.

La spiegazione basata su esempi spesso ottimizza gli strati nascosti in profondità della rete neurale (cioè lo spazio latente) in modo tale che i punti simili siano vicini tra loro in questo spazio latente, mentre i punti dissimili siano più lontani nello spazio latente.

Nella Tabella 4 è riportata una panoramica dei lavori che utilizzano la spiegazione basata su esempi nell'imaging medico.

Posizione anatomica	Autori (anno)	Modalità	Tecnica XAI utilizzata/basata su
Cervello	Li et al. 2019d	RISONANZA MAGNETICA	Esempi dallo spazio latente
Seno	Uehara et al. (2019)	Istologia	Prototipi
Petto	LaLonde et al. (2020)	CT	Esempi dallo spazio latente
	Silva et al. (2020)	Raggi X	Esempi dallo spazio latente
Gastrointestinale	Peng et al. (2019)	Istologia	Rete di triplette
	Wang et al. (2019)	RISONANZA MAGNETICA	Funzioni di influenza
La pelle	Codella et al. (2018)	Dermatoscopia	Rete di triplette
	Sarhan et al. (2019)	Dermatoscopia	Esempi dallo spazio latente
Tiroide	Chen et al. (2020)	Istologia	Esempi dallo spazio latente
	Li et al. 2020	Ultrasuoni	Prototipi
Multiplo	Biffi et al. (2020)	RISONANZA MAGNETICA	Esempi dallo spazio latente
	Choudhary et al. (2019)	Istologia	Rete di triplette
	Silva et al. (2018)	Multiplo	Esempi dallo spazio latente
	Yan et al. (2018)	CT	Rete di triplette
	Yang et al. (2020)	Istologia	Esempi dello spazio latente con spiegazione visiva

Tabella 4. Papers che forniscono spiegazioni basate su esempi. Per motivi di leggibilità, i papers sono ordinati in base alla posizione anatomica e solo il primo paper che tratta tale posizione anatomica riporta il nome della posizione. La colonna "Principale tecnica XAI utilizzata/basata su" descrive quale tecnica di spiegazione basata su esempi è stata utilizzata nella Sezione 3.3, o su quale tecnica si basa il metodo del paper corrispondente. TC = tomografia computerizzata, RM = risonanza magnetica.

3.3.1 Rete di triplette

Diversi lavori hanno fornito spiegazioni basate su esempi utilizzando una rete tripla (Hoffer e Ailon, 2015). Una rete tripla consiste in tre reti identiche con parametri condivisi. Alimentando queste reti con tre campioni in ingresso, la rete calcola due valori che consistono nelle distanze *L2* tra le rappresentazioni nello spazio latente (cioè le rappresentazioni incorporate) di questi campioni in ingresso. Ciò consente di apprendere le rappresentazioni utili attraverso un confronto non supervisionato dei campioni. Quando si analizza un punto di dati, l'ispezione dei vicini in questa rappresentazione incorporata fornirà esempi di punti di dati simili al punto di dati che si sta analizzando, che possono fornire una spiegazione del perché la rete è arrivata al suo output.

Peng et al. (2019) hanno utilizzato una spiegazione basata su esempi nell'istologia del cancro colorettale. Per prima cosa hanno addestrato una CNN utilizzando una perdita tripla, hashing e *k* hard-negative per apprendere un embedding che preserva la somiglianza. In fase di test, una ricerca da grossolana a fine ha prodotto i 10 esempi più vicini da un database di test relativi all'immagine di input. In questo modo è stato spiegato su quali immagini simili all'immagine analizzata la rete ha basato la propria decisione.

Yan et al. (2018) hanno utilizzato un sistema di archiviazione e comunicazione di immagini radiologiche (PACS) per estrarre 32000 lesioni clinicamente rilevanti dall'intero corpo. Per apprendere le incorporazioni rilevanti delle lesioni, hanno addestrato una rete tripla con tre spunti di supervisione: dimensione della lesione, posizione anatomica della lesione (ad esempio, polmone, fegato o rene) e coordinata relativa della lesione nel corpo. Queste incorporazioni hanno mostrato una buona separazione in base alla posizione anatomica (ad esempio, le lesioni epatiche sono state separate da quelle polmonari) e sono state in grado di recuperare accuratamente le spiegazioni basate su esempi da un set di test.

Anche Codella et al. (2018) hanno utilizzato una perdita di triplette, ma l'hanno combinata con il pooling della media globale, la tecnica utilizzata in CAM. Di conseguenza, non solo hanno potuto estrarre spiegazioni basate su esempi, ma hanno anche fornito mappe di attivazione della query e mappe di attivazione dei risultati della ricerca. In altre parole, una spiegazione visiva mostrava quale regione dell'immagine di input era stata utilizzata dalla rete per generare la spiegazione basata sull'esempio. Hanno dimostrato questa tecnica su immagini dermatologiche di melanoma.

3.3.2 Funzioni di influenza

Wei Koh e Liang (2017) hanno proposto di utilizzare le funzioni di influenza per spiegare su quali input di un set di formazione si basa una decisione. Lo hanno fatto studiando cosa accadrebbe nel caso in cui un input dell'insieme di addestramento non fosse disponibile o venisse modificato. Poiché è costoso valutare questo aspetto mediante perturbazione, hanno fornito un'approssimazione efficiente utilizzando le funzioni di influenza (Cook e Weisberg, 1980). Questa implementazione delle funzioni di influenza è correlata a SHAP, nel senso che entrambe consentono di calcolare in modo efficiente l'importanza delle caratteristiche.

Wang et al. (2019) hanno utilizzato funzioni di influenza per spiegare quali classificazioni di lesioni epatiche su RM multifase erano associate a quali caratteristiche radiologiche. Questa spiegazione globale ha fornito indicazioni sul comportamento della rete neurale. Ad esempio, la classe "cisti benigna" era più spesso associata al reperto radiologico "massa a pareti sottili". Poiché la rete non forniva solo l'etichetta della classe, ma anche le caratteristiche radiologiche corrispondenti, questa spiegazione poteva aumentare la fiducia dell'utente nei risultati della rete.

3.3.3 Prototipi

Chen et al. 2019 hanno proposto di utilizzare esempi tipici come spiegazione (cioè prototipi), che hanno descritto come "questo-sembra-così". Il metodo riflette il ragionamento basato

sui casi che gli esseri umani eseguono. Ad esempio, quando una persona spiega perché un'immagine contiene un'auto, può internamente ragionare sul fatto che si tratta di un'auto perché assomiglia a un'auto che ha già visto. Alla rete neurale è stato aggiunto uno strato di prototipi, che raggruppa gli input di addestramento in base alle loro classi nello spazio latente. Per ogni classe è stato scelto un prototipo, costituito da un esempio tipico di quella classe. Durante il test, il metodo utilizzava parti dell'immagine di prova che assomigliavano a questi prototipi addestrati. L'output era una combinazione ponderata delle somiglianze con questi prototipi. Pertanto, la spiegazione era un calcolo effettivo della rete neurale, non un'approssimazione post hoc.

Uehara et al. (2019) hanno utilizzato i prototipi per spiegare perché una rete neurale ha classificato patch di immagini istologiche come cancro o non cancro. La rete è stata in grado di identificare su quali parti dell'immagine ha basato la sua decisione e in che misura queste parti dell'immagine erano simili a esempi prototipici appresi dal set di addestramento.

[...]

4. PRO E CONTRO DELLE TECNICHE XAI

Tutte le tecniche XAI descritte nella Sezione 3 hanno pro e contro che influenzano la scelta tra le varie opzioni. I pro e i contro saranno strutturati nelle categorie facilità d'uso, validità, robustezza, costo computazionale, necessità di messa a punto e disponibilità di risorse aperte. Una panoramica di questi pro e contro per ogni metodo della Tabella 1 è riportata nella Tabella 5.

Tecnica	Facilità d'uso	Validità	Robustezza	Esigenze di calcolo	Non è necessaria alcuna messa a punto	Open-source (paper originale)	Open-source (captum.ai)
Spiegazione visiva							
Approcci basati sulla retropropagazione							

Retropropagazione	+	-	+	–	+	–	+
Deconvoluzione	+	n.t.	n.t.	–	+	–	+
Retropropagazione guidata	+	-	inc.	–	+	–	+
Mappatura dell'attivazione della classe (CAM)	+	n.t.	–	–	+	a	–
Mappatura dell'attivazione della classe pesata con gradiente (Grad-CAM)	+	+/-	–	–	+/–	b	+
Propagazione della rilevanza a livello di strato (LRP)	+	n.t.	+	–	+/–	–	+
Esopiani additivi Deep SHapley (Deep SHAP)	+	n.t.	n.t.	–	+/-	c	+
Attenzione allenabile	+/-	n.t.	n.t.	+	–	d	–
Approcci basati sulla perturbazione							
Sensibilità all'occlusione	+	n.t.	–	+	–	–	+
Spiegazioni locali interpretabili e modello-agnostico (LIME)	+	n.t.	n.t.	+	–	e	+
Perturbazione significativa	+	n.t.	n.t.	+	–	f	–
Analisi delle differenze di previsione	+	n.t.	n.t.	+	–	g	–
Spiegazione testuale							
Didascalia dell'immagine	+/–	n.t.	n.t.	+	–	–	–
Didascalia delle immagini con spiegazione visiva	+/–	n.t.	n.t.	+	–	h	–
Test con vettori di attivazione del concetto (TCAV)	+	n.t.	n.t.	n.t.	+/–	i	–

Spiegazione basata su esempi							
Reti di triplette	+/–	n.t.	n.t.	+	–	j	–
Funzioni di influenza	+	n.t.	n.t.	n.t.	+/–	k	–
Prototipi	+/–	n.t.	n.t.	+	–	l	–

Tabella 5. Pro e contro delle tecniche XAI. I pro sono rappresentati da +, i contro da –. Le lettere nella colonna Open source (paper originale) si riferiscono all'URL sotto la tabella.

- *n.t. = non testato da studi su quel criterio.*
- *inc. = risultati non conclusivi tra gli studi su quel criterio.*
- https://github.com/zhoubolei/CAM
- *b* https://github.com/Cloud-CV/Grad-CAM
- *c* https://github.com/slundberg/shap
- *d* https://github.com/saumya-jetley/cd_ICLR18_LearnToPayAttention
- *e* https://github.com/marcotcr/lime
- *f* https://github.com/ruthcfong/perturb_explanations
- *g* https://github.com/lmzintgraf/DeepVis-PredDiff
- *h* https://github.com/zizhaozhang/tandemnet
- *i* https://github.com/tensorflow/tcav
- *j* https://github.com/eladhoffer/TripletNet
- *k* https://github.com/kohpangwei/influence-release
- *l* https://github.com/cfchen-duke/ProtoPNet

[...]

4.6 DISPONIBILITÀ OPEN-SOURCE

La maggior parte delle tecniche XAI sono disponibili in open source. Spesso il codice è disponibile presso gli autori dell'articolo originale. Molte tecniche sono anche implementate in pacchetti XAI come captum.ai. Una panoramica della disponibilità di tecniche XAI open source è riportata nella Tabella 5.

5. DISCUSSIONE

5.1 PANORAMICA

Abbiamo discusso 223 articoli sull'intelligenza artificiale spiegabile (XAI) per il deep learning nell'analisi delle immagini

mediche. Abbiamo classificato gli articoli in base alle strutture XAI proposte da Adadi e Berrada (2018) e Murdoch et al. (2019). Alcune tendenze sono state notate negli articoli esaminati. La maggior parte dei lavori ha utilizzato spiegazioni post hoc rispetto a quelle basate su modelli, ossia le spiegazioni sono state fornite su una rete neurale già addestrata, invece di essere incorporate nell'addestramento della rete neurale. Sono stati utilizzati metodi di spiegazione sia specifici per il modello (ad esempio, progettati specificamente per le CNN) che indipendenti dal modello. Inoltre, la maggior parte dei lavori esaminati ha fornito spiegazioni locali piuttosto che globali, cioè la spiegazione è stata fornita per ogni caso (ad esempio, per ogni paziente), piuttosto che a livello di set di dati (ad esempio, per tutti i pazienti). Poiché ci concentriamo sul deep learning nell'analisi delle immagini mediche, queste tendenze erano prevedibili. La maggior parte dei metodi XAI disponibili, adatti alle CNN, sono tecniche di mappatura della salienza, che spesso forniscono spiegazioni post hoc, specifiche per il modello e locali. Inoltre, i metodi XAI post hoc possono essere utilizzati dopo l'addestramento di una rete neurale, rendendoli più accessibili rispetto a quelli basati sul modello.

Abbiamo classificato gli articoli in base alla posizione anatomica e alla modalità di imaging medico. Abbiamo riscontrato che la maggior parte degli articoli si concentra sul torace o sul cervello e sulla risonanza magnetica (Fig. 3). Ciò è paragonabile a quanto riscontrato da Litjens et al. (2017) per i metodi di deep learning nell'imaging medico in generale. Questa tendenza è probabilmente dovuta ai set di dati disponibili pubblicamente in questi organi e modalità e non riflette la capacità di spiegazione di questi organi e modalità.

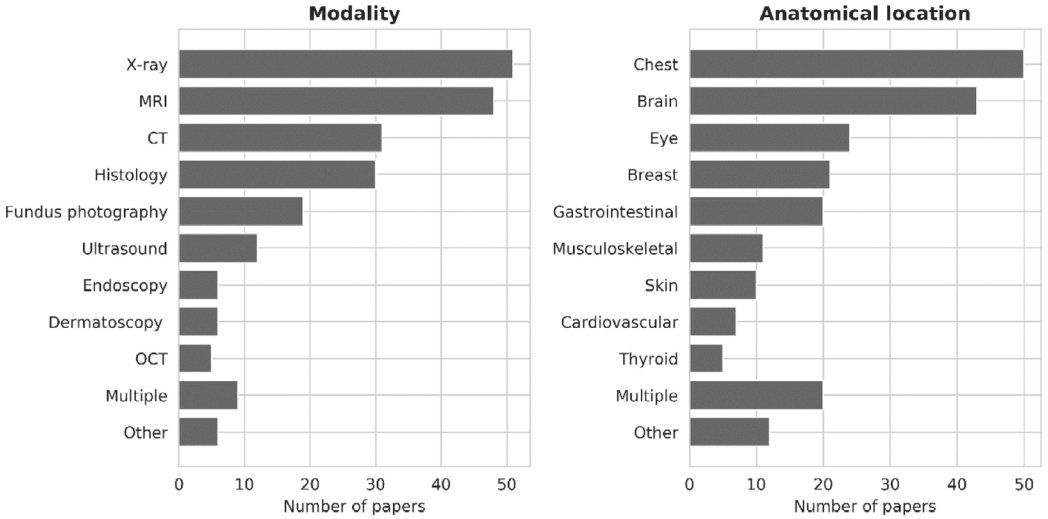


Fig. 3. Papers inclusi in questa indagine, classificati per modalità (a sinistra) e sede anatomica (a destra). Gli articoli che discutevano più modalità o sedi anatomiche sono stati raggruppati come "multipli". Le modalità o le sedi anatomiche utilizzate in meno di cinque articoli sono state raggruppate come "altro".

5.2 VALUTAZIONE DI XAI

Abbiamo descritto diverse tecniche XAI e le loro applicazioni nell'analisi delle immagini mediche, ma come si valuta se una tecnica XAI fornisce una buona spiegazione? A differenza delle misure di performance comunemente utilizzate nell'analisi delle immagini mediche, come l'accuratezza, il coefficiente Dice o l'analisi ROC, i criteri di successo della spiegazione sono più difficili da definire. Doshi-Velez e Kim (2017) hanno proposto un quadro per la valutazione della spiegabilità, composto da tre metodi di valutazione: valutazione basata sull'applicazione, valutazione basata sull'uomo e valutazione basata sulla funzione.

5.2.1 Valutazione basata sull'applicazione

La valutazione basata sull'applicazione utilizza esperimenti umani all'interno di un'applicazione reale. In altre parole, si la-

scia che siano gli esperti del dominio a testare la spiegazione. Nell'analisi delle immagini mediche, ciò potrebbe comportare che un radiologo verifichi se le spiegazioni basate su esempi sono effettivamente buoni esempi, sulla base delle molte immagini che il radiologo ha visto nei suoi molti anni di esperienza. Il vantaggio della valutazione basata sull'applicazione è che verifica direttamente l'obiettivo per cui il sistema è stato costruito. Lo svantaggio è che si tratta di una valutazione costosa.

5.2.2 Valutazione basata sull'uomo

La valutazione human-grounded utilizza esperimenti umani più semplici che mantengono l'essenza dell'applicazione target. In altre parole, si lascia che siano i non addetti ai lavori a testare la spiegazione o un proxy della spiegazione. Ad esempio, per spiegare la posizione e le dimensioni di un tumore, si potrebbe ricorrere a un progetto di crowdsourcing in cui i non addetti ai lavori giudicano la qualità delle mappe di salienza. Poiché utilizza persone non esperte invece di esperti di dominio altamente qualificati, il vantaggio della valutazione basata sull'uomo è che è meno costosa, pur ricevendo nozioni generali sulla qualità di una spiegazione. Lo svantaggio è che la valutazione della qualità di una spiegazione è una proxy della qualità effettiva.

5.2.3 Valutazione fondata sulle funzioni

La valutazione con base funzionale non utilizza esperimenti sull'uomo, ma si avvale di altre proxy per valutare la qualità della spiegazione. Queste proxy possono includere misure che sono già state convalidate da utenti umani. Nel nostro esempio di spiegazione della posizione e delle dimensioni di un tumore, ciò potrebbe comportare il confronto della spiegazione con le delineazioni del tumore tracciate manualmente da un radiologo. I vantaggi della valutazione con base funzionale dichiarati da Doshi-Velez e Kim (2017) includono il fatto che sono relativamente economici da acquisire. Tuttavia, questo non è necessariamente il caso dell'analisi delle immagini mediche, poiché l'acquisizione

di annotazioni manuali, ad esempio, è un processo che richiede molte risorse. Quando queste annotazioni manuali esistono già, ad esempio quando si utilizzano i dati curati di una sfida, la valutazione delle spiegazioni è facilmente estraibile e può essere estratta automaticamente più volte. Questo può essere utile, ad esempio, nella fase di sviluppo dei metodi di spiegazione.

5.2.4 Valutazione di XAI nell'analisi delle immagini mediche

La valutazione della XAI proposta sopra non è ancora una pratica standard nei papers sull'analisi delle immagini mediche. Inoltre, in medicina una buona spiegazione può variare a seconda delle aree di competenza della persona a cui viene fornita. Ad esempio, una spiegazione visiva che indichi la posizione della malattia potrebbe essere sufficiente per un radiologo o un ricercatore di analisi delle immagini mediche. Tuttavia, i medici come gli oncologi, i neurologi o gli ematologi probabilmente vorrebbero che la XAI fosse aggiunta al loro quadro decisionale clinico. Tale quadro includerebbe anche l'anamnesi del paziente, i trattamenti precedenti e attuali, le opzioni terapeutiche e gli effetti o i risultati attesi.

5.3 CRITICA SU XAI

Rudin (2019) ha consigliato cautela nell'utilizzo di una scatola nera con spiegazione per i processi decisionali ad alto rischio. Rudin ha sollevato diversi problemi relativi alla spiegazione delle scatole nere. Ad esempio, XAI può fornire una spiegazione che non è completamente fedele a ciò che il modello originale calcola: Se la spiegazione è fedele al modello per il 90%, significa che il 10% è falso (Rudin, 2019). Inoltre, una spiegazione potrebbe non avere senso o non fornire abbastanza dettagli per capire cosa sta facendo la scatola nera. Ad esempio, una mappa di salienza della classe con la probabilità più alta può sembrare simile a una mappa di salienza di una classe con una probabilità più bassa. Rudin consiglia quindi di utilizzare una XAI basata su un modello interpretabile, come il prototipo di rete discusso nella sezione

3.3.3. Le critiche si concentrano spesso anche sulla robustezza delle tecniche XAI, come discusso nella Sezione 4.

5.4 PROSPETTIVE

Poiché il processo decisionale ad alto rischio è intrecciato con la medicina, siamo convinti che la XAI sarà sempre più importante. Abbiamo analizzato le tendenze in atto e abbiamo notato che un numero crescente di articoli contiene un approccio olistico, che combina più forme di spiegazione. Esempi di questi approcci più olistici sono le combinazioni di spiegazioni testuali e visive (ad esempio, Graziani et al., 2020), o le combinazioni di spiegazioni basate su esempi e spiegazioni visive (ad esempio, Wang et al., 2019).

Le direzioni future della XAI nell'analisi delle immagini mediche potrebbero includere la spiegazione biologica. Diversi ricercatori hanno previsto i processi biologici dalle caratteristiche delle immagini utilizzando il deep learning. Ad esempio, Matsui et al. (2020) hanno previsto il sottotipo molecolare dei gliomi di grado inferiore sulla base di immagini cerebrali multimodali e Zhu et al. (2019) hanno previsto il sottotipo molecolare luminale A del cancro al seno sulla base della risonanza magnetica. Queste analisi hanno utilizzato un target biologico per addestrare la rete neurale. Tuttavia, l'esecuzione di tali analisi al contrario, ad esempio eseguendo un'analisi dei percorsi sui fenotipi di imaging (ad esempio Bismeijer et al. (2020), non deep learning), potrebbe fornire interessanti spiegazioni biologiche.

Le XAI possono anche essere utili per aiutare i medici nel processo diagnostico o nell'identificazione di informazioni sconosciute dalle immagini mediche. Ad esempio, uno studio sulla diagnosi di tubercolosi su radiografie del torace ha dimostrato che 10 dei 13 medici partecipanti (77%) hanno ottenuto una migliore accuratezza diagnostica quando hanno valutato le radiografie del torace con un XAI che forniva una spiegazione visiva rispetto alla valutazione della radiografia del torace senza XAI (Rajpurkar et al., 2020a).

È probabile che la XAI nell'imaging medico includa sempre più informazioni di dominio. Per raggiungere questo obiettivo, i medici dovrebbero essere coinvolti nella progettazione di metodi di interpretazione specifici per il compito (Fan et al., 2021). La collaborazione attiva tra medici, ricercatori teorici, esperti di imaging medico e di analisi delle immagini mediche sarà una strada importante per lo sviluppo futuro dei metodi di deep learning (Fan et al., 2021).

Altre direzioni della XAI nell'analisi delle immagini mediche possono includere il legame tra causalità e XAI. La tipica analisi delle immagini mediche consiste nella correlazione piuttosto che nella causalità. La causalità descrive la relazione tra causa ed effetto e può essere descritta matematicamente (Pearl, 2009). Le attuali tecniche XAI che mirano a essere prive di pregiudizi, come i prototipi, sono potenzialmente ancora sensibili alle differenze nella popolazione di addestramento, il che potrebbe ostacolare la generalizzabilità. Castro et al. (2020) descrivono come il ragionamento causale possa essere utile per valutare i bias nei dati. DeGrave et al. (2021) hanno fornito un esempio di come i bias dei dataset possano essere rilevati utilizzando l'XAI: negli studi che distinguono tra le radiografie di pazienti positivi al Coronavirus 2019 (COVID-19) e quelle di pazienti negativi al COVID-19, hanno utilizzato una spiegazione visiva per dimostrare che le elevate prestazioni dei modelli di deep learning erano in realtà attribuite al modo in cui erano stati composti i dataset, piuttosto che all'effettivo rilevamento del COVID-19 nelle radiografie. van Amsterdam et al. (2019) mostrano un esempio di eliminazione dei pregiudizi utilizzando la causalità, ottenendo una previsione imparziale della prognosi per i pazienti affetti da cancro ai polmoni. Sarebbe interessante incorporare tali analisi nella spiegazione delle immagini mediche, come hanno fatto Chattopadhyay et al. (2019) per la spiegazione visiva dei dati MNIST.

Non esiste un consenso sulle stime a priori delle dimensioni del campione necessarie per XAI e deep learning nell'imaging medico in generale (Balki et al., 2019). Data la natura costosa dell'acquisizione di set di dati di imaging medico in termini di

denaro, tempo e onere per il paziente, è auspicabile disporre di linee guida che descrivano quali dimensioni minime del campione sarebbero necessarie per quali tecniche XAI.

[...]

7 - APPLICAZIONI DELL'INTELLIGENZA ARTIFICIALE SPIEGABILE NELLA DIAGNOSI E NELLA CHIRURGIA

Tratto e tradotto da
Yiming Zhang, Ying Weng e Jonathan Lund (2022). Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. In (Ed.) Diagnostics, 2022, *12*(2), 237.

https://doi.org/10.3390/diagnostics12020237

Le parti omesse dal curatore rispetto all'originale sono indicate dal segno **[...]**

ABSTRACT

Negli ultimi anni, l'intelligenza artificiale (IA) ha mostrato grandi promesse in medicina. Tuttavia, i problemi di spiegabilità rendono difficile l'applicazione dell'IA nell'uso clinico. Sono state condotte alcune ricerche sull'intelligenza artificiale spiegabile (XAI, Explainable Artificial Intelligence) per superare la limitazione della natura black-box dei metodi di IA. Rispetto alle tecniche di IA come il deep learning, la XAI può fornire sia il processo decisionale che le spiegazioni del modello. In questa revisione, abbiamo condotto un'indagine sulle recenti tendenze nella diagnosi medica e nelle applicazioni chirurgiche che utilizzano XAI. Abbiamo cercato articoli pubblicati tra il 2019 e il 2021 su PubMed, IEEE Xplore, Association for Computing Machinery e Google Scholar. Abbiamo incluso gli articoli che soddisfacevano i criteri di selezione nella revisione e poi abbiamo estratto e analizzato le informazioni rilevanti dagli studi. Inoltre, forniamo una vetrina sperimentale sulla diagnosi del cancro al

seno e illustriamo come XAI possa essere applicato nelle applicazioni mediche XAI. Infine, riassumiamo i metodi XAI utilizzati nelle applicazioni mediche XAI, le sfide che i ricercatori hanno affrontato e discutiamo le direzioni di ricerca future. I risultati dell'indagine indicano che la XAI medica è una direzione di ricerca promettente e questo studio vuole essere un riferimento per gli esperti di medicina e gli scienziati dell'IA nella progettazione di applicazioni XAI mediche.

1. INTRODUZIONE

Il Machine Learning (ML) e il Deep Learning (DL) hanno compiuto progressi impressionanti negli ultimi tempi e il successo dell'intelligenza artificiale (IA) in campo medico ha portato a un aumento significativo delle applicazioni di IA in campo medico. L'obiettivo della ricerca sull'IA medica è quello di creare applicazioni che utilizzino le tecnologie dell'IA per assistere i medici nelle decisioni mediche [1]. L'IA viene utilizzata in diverse applicazioni mediche, come la diagnosi delle malattie [2], la chirurgia [3] e molte altre. Tuttavia, le applicazioni di IA in campo medico devono affrontare alcune sfide, tra cui la natura black-box di alcuni modelli di IA. La scarsa spiegabilità di questi modelli black-box porta alla sfiducia degli esperti medici nel fare inferenze cliniche spiegabili.

Spesso i modelli DL contengono milioni di parametri e restituiscono solo un risultato finale senza alcuna spiegazione. A causa della mancanza di trasparenza delle reti neurali profonde, è difficile per l'utente giudicare se la decisione è affidabile, compromettendo la fiducia dei medici. Le applicazioni di IA medica devono essere trasparenti per aumentare il livello di fiducia dei medici. La ricerca sull'intelligenza artificiale spiegabile (XAI) ha recentemente guadagnato una notevole attenzione [4]. Affinché le applicazioni di IA medica siano accettate e integrate nella pratica, la XAI è fondamentale.

1.1 CONCETTI DI INTELLIGENZA ARTIFICIALE CORRELATI

L'intelligenza artificiale si riferisce allo sviluppo dell'intelligenza da parte delle macchine e l'apprendimento automatico fa parte dell'IA. Un algoritmo di ML viene addestrato fornendo molti esempi per un determinato compito, trovando il modello statistico in questi esempi e, infine, scoprendo le regole per automatizzare il compito [5]. Gli algoritmi di ML tradizionali, tra cui k-nearest neighbor (kNN), support vector machine (SVM), decision tree (DT) e random forest (RF), sono stati applicati nell'IA medica. kNN è un algoritmo che trova i punti di dati più vicini nel set di addestramento da utilizzare come predizione per i nuovi dati [6]. SVM presuppone che i dati siano linearmente separabili e cerca di trovare un iperpiano lineare (confine decisionale) per separare i dati. Gli esempi in un modello SVM sono rappresentati come punti nello spazio, divisi in categorie separate da un iperpiano lineare [7]. Un albero decisionale è una struttura ad albero in cui ogni nodo interno rappresenta un test di attributo, ogni ramo rappresenta un risultato del test e ogni foglia indica la classe. Nel DT, l'idea di base è quella di scomporre una decisione complessa in diverse decisioni più semplici, in modo che il risultato finale assomigli al risultato desiderato [8]. RF è un algoritmo di apprendimento automatico di tipo ensemble che consiste in molti alberi decisionali. Per i compiti di classificazione, la decisione di RF è il risultato della votazione di questi alberi decisionali [9].

Oltre ai metodi tradizionali di apprendimento automatico, molti studi hanno utilizzato anche metodi di deep learning per applicazioni mediche. Un algoritmo di deep learning è in grado di apprendere rappresentazioni di dati grezzi senza l'ingegneria delle caratteristiche. I tipici metodi di deep learning includono i perceptron multistrato (MLP), le reti neurali profonde (DNN), le reti neurali convoluzionali (CNN) e le reti neurali ricorrenti (RNN) [10]. Inoltre, in termini di valutazione delle prestazioni dei metodi di ML, le metriche di valutazione tipiche sono l'ac-

curatezza, la precisione, il richiamo, il punteggio F1, l'AUC e il ROC [11].

1.2 CONCETTI DI INTELLIGENZA ARTIFICIALE SPIEGABILI CORRELATI

Secondo [12], la spiegabilità è la capacità di spiegare il processo decisionale dell'IA in termini comprensibili per gli esseri umani, con una gamma più ampia di utenti finali su come è stata presa una decisione. I diversi utenti finali si concentrano su prospettive diverse di spiegabilità. Gli esperti di IA o i data scientist sono più interessati alla spiegabilità del modello/algoritmo. Tuttavia, gli esperti di medicina o i medici sono più interessati all'inferenza/predizione clinica. L'altra nozione correlata alla spiegabilità è l'interpretabilità. Per interpretabilità si intende la capacità di fornire il significato di un concetto astratto [13]. La spiegabilità si riferisce all'interpretazione delle previsioni fatte in presenza di nuovi casi, mentre l'interpretabilità si riferisce alla resa del modello appreso dai dati durante l'addestramento [14]. Inoltre, esistono due tipi di metodi XAI: intrinseci e post hoc [15]. Un metodo intrinseco è quello che ci permette di comprendere un processo decisionale o la base di una tecnica senza informazioni aggiuntive. I metodi intrinseci tipici includono la regressione lineare (LR) [16], la regressione logistica, il k-nearest neighbor, gli apprendisti basati su regole, i modelli additivi generali, i modelli bayesiani e gli alberi decisionali. L'apprendimento profondo è un sottoinsieme dell'apprendimento automatico e l'apprendimento automatico è un sottoinsieme dell'IA. Inoltre, riteniamo che XAI sia un sottoinsieme dell'IA e che i suoi metodi intrinseci siano ML. Pertanto, nella Figura 1, mostriamo la relazione tra intelligenza artificiale, machine learning, deep learning e intelligenza artificiale spiegabile.

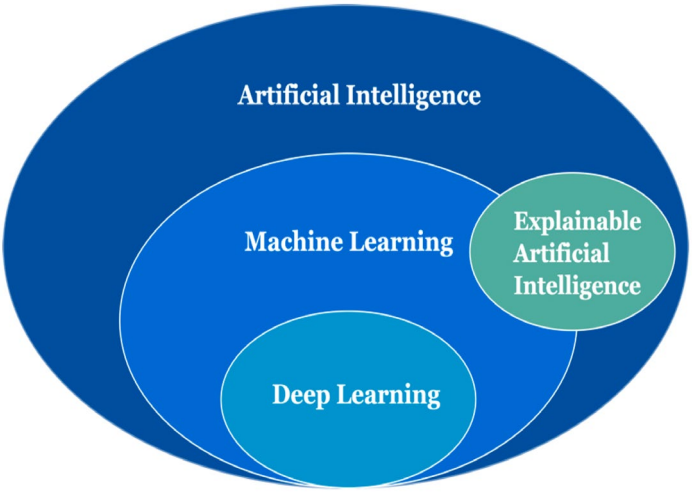


Figura 1. Il rapporto tra intelligenza artificiale, apprendimento automatico, apprendimento profondo e intelligenza artificiale spiegabile.

Utilizzando questo metodo, è possibile capire quale parte dei dati di input è responsabile della decisione di classificazione in qualsiasi classificatore. Altri metodi post hoc includono SHapley Additive exPlanations (SHAP) [17], class activation mapping (CAM) [18], principle component analysis (PCA) [19] e Gradient-weighted class activation mapping (Grad-CAM) [20]. Secondo [21,22], i metodi di spiegabilità post hoc possono essere classificati come: riduzione delle dimensioni, meccanismo di attenzione, architettura di rete neurale ristretta, spiegazione testuale, spiegazione visiva, spiegazione locale, spiegazione per esempio, spiegazione per semplificazione e rilevanza delle caratteristiche. La tassonomia dei metodi XAI è illustrata nella Figura 2.

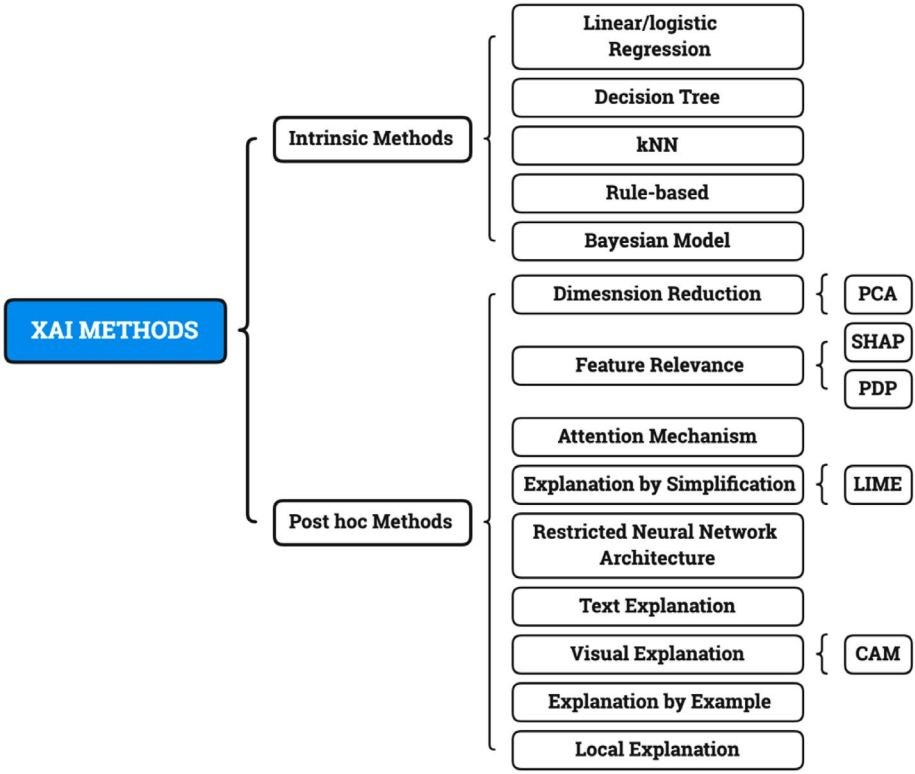


Figura 2. Tassonomia dei metodi XAI, tipi di XAI post hoc e alcuni esempi.

Per la valutazione delle XAI non è stata adottata alcuna metrica di valutazione oggettiva o unificata. Doshi-Velez e Kim definiscono tre tipi di approcci alla valutazione XAI [12]: valutazioni basate sull'applicazione, valutazioni basate sull'uomo e valutazioni basate sulla funzione.

1.3 CONTRIBUTI

Ci sono stati pochi studi che hanno esplorato il potenziale di XAI nelle applicazioni di IA in campo medico [23]. Questo

studio si concentra sulle applicazioni XAI in campo medico per la diagnosi e la chirurgia. La Figura 3 mostra la pipeline generale di un'applicazione XAI in campo medico. Come mostrato nella Figura 3, l'utilizzo del metodo XAI intrinseco consente all'applicazione XAI medica di esaminare i dati medici e di fornire decisioni e spiegazioni ai medici. In alternativa, se l'applicazione medica utilizzasse uno XAI post hoc, i metodi black-box verrebbero applicati ai dati medici per prendere decisioni, seguiti dallo XAI post hoc che fornisce una spiegazione dei metodi black-box.

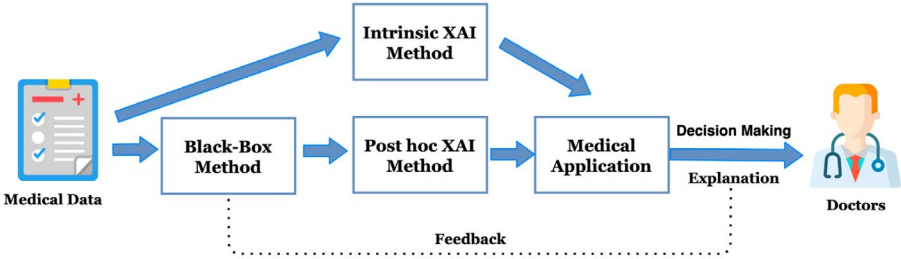


Figura 3. La pipeline complessiva di un'applicazione XAI medica: i metodi XAI possono essere intrinseci o post hoc e possono fornire decisioni e spiegazioni ai medici.

Negli ultimi anni, l'importanza di XAI è stata ampiamente riconosciuta nel mondo accademico e industriale. A causa dell'elevato grado di complessità, il processo decisionale del modello di deep learning è difficile da interpretare. Inoltre, la natura black-box di questi modelli è pericolosa se vengono impiegati in applicazioni cliniche, perché potrebbero non fornire giustificazioni affidabili agli esperti medici. Molti studi XAI sono stati proposti nella comunità dell'IA per superare questo problema. Tuttavia, nel campo interdisciplinare dell'intelligenza artificiale e della medicina, i modelli di deep learning sono la scelta principale per la maggior parte delle applicazioni di IA medica. Pertanto, è essenziale utilizzare e sviluppare metodi XAI invece di metodi black-box di deep learning. Abbiamo riscontrato che la maggior parte delle indagini sulle applicazioni di IA in campo medico utilizza solo il deep learning, ma non c'è stata alcuna in-

dagine incentrata sulle applicazioni di IA in campo medico che utilizzano l'XAI, in particolare per la diagnosi e la chirurgia. Riteniamo che un'indagine di questo tipo possa fornire agli esperti di medicina e di IA informazioni sui recenti progressi delle applicazioni mediche XAI. Inoltre, sarà utile ai ricercatori di medicina e di IA per sviluppare le loro applicazioni XAI in campo medico. Questa indagine mira a rispondere alle seguenti tre domande di ricerca (RQ): (RQ1) Quali sono le attuali tendenze della ricerca sulle applicazioni XAI in campo medico? (RQ2) In che modo gli studi inclusi in questa indagine affrontano il compromesso tra accuratezza e spiegabilità? e (RQ3) È possibile implementare questi modelli nell'ambiente clinico del mondo reale per assistere gli esperti medici e fare un'inferenza clinica spiegabile? In sintesi, i principali contributi di questa indagine includono:

- Una breve introduzione ai concetti di IA/DL, ai concetti di XAI e alla pipeline generale delle applicazioni mediche XAI fornisce un rapido inizio per gli esperti del settore medico;
- La nostra indagine fornisce anche una recente rassegna triennale della letteratura generale sulle applicazioni XAI in campo medico nei settori della diagnosi e della chirurgia, con un'analisi approfondita;
- Riassumiamo le tendenze attuali e discutiamo le sfide e le direzioni future su come progettare una migliore applicazione XAI medica.

Il resto dell'articolo è strutturato come segue: La Sezione 2 descrive la strategia di ricerca dell'indagine; la Sezione 3 presenta i risultati della selezione delle applicazioni XAI mediche per la diagnosi e la chirurgia; nella Sezione 4, presentiamo la discussione dell'indagine, compresi i risultati dello studio, la vetrina sperimentale, le sfide, i limiti, le lacune della ricerca, nonché le direzioni future e le risposte alle domande di ricerca; la Sezione 5 conclude l'indagine.

2. STRATEGIA DI RICERCA

È stata condotta una ricerca in letteratura utilizzando le parole chiave intelligenza artificiale spiegabile, diagnostica e chirurgia. Inoltre, gli articoli di ricerca citati in questa revisione sono stati trovati su tre database elettrotonici - PubMed, IEEE Xplore e Association for Computing Machinery (ACM) - per le pubblicazioni pertinenti tra il 2019 e il 2021 incluso. Anche Google Scholar è stato cercato tra queste date per studi potenzialmente rilevanti. Le stringhe di ricerca utilizzate in questa indagine: (ALL("Explainable Artificial Intelligence") OR ALL("XAI") OR ALL("Explainable AI") OR ALL("Diagnostics") OR ALL("Surgery") OR ALL("Medical")). L'indagine mira a individuare le pubblicazioni sulle applicazioni XAI in campo medico nella diagnostica e nella chirurgia. Pertanto, tutti gli articoli inclusi dovevano essere incentrati su questo argomento. Sono stati esclusi tutti gli articoli di indagine/revisione, gli articoli non in inglese o non sottoposti a revisione paritaria. I titoli e gli abstract di questi articoli di ricerca sono stati quindi esaminati per verificarne l'ammissibilità. Successivamente, tutti gli articoli di ricerca selezionati sono stati esaminati per verificarne la rilevanza nel testo completo; gli articoli idonei sono stati inclusi in questa revisione.

3. APPLICAZIONI DI INTELLIGENZA ARTIFICIALE SPIEGABILI IN CAMPO MEDICO

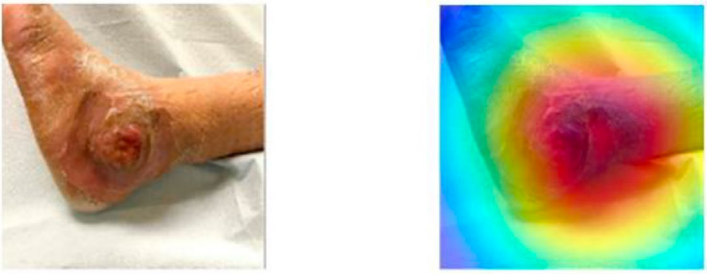
3.1 DIAGNOSI

In [24], Kavya et al. hanno proposto un framework assistito da computer per la diagnosi delle allergie. Gli autori hanno applicato diversi algoritmi di ML e hanno poi selezionato l'algoritmo con le migliori prestazioni utilizzando la convalida incrociata k-fold. Per quanto riguarda il metodo XAI, hanno sviluppato un approccio basato su regole costruendo una foresta casuale. In

particolare, ogni percorso in un albero è rappresentato come una regola IF-THEN e le spiegazioni sono estratte dai dati medici. Inoltre, gli autori hanno implementato il framework computerizzato su un'applicazione mobile, che può aiutare i medici junior a confermare le previsioni diagnostiche. In [25], Amoroso et al. hanno presentato un framework XAI per le terapie del cancro al seno. Hanno applicato il metodo di clustering e di riduzione delle dimensioni e i risultati degli esperimenti hanno dimostrato che il framework è in grado di delineare le caratteristiche cliniche più importanti per il paziente e di progettare terapie oncologiche. Dindorf et al. hanno proposto un classificatore indipendente dalla patologia spiegabile per la postura spinale [26]. Gli autori hanno utilizzato SVM e RF come classificatori ML e hanno poi applicato LIME per spiegare la previsione del classificatore ML. In [27], El-Sappagh et al. hanno proposto un modello RF per la diagnosi e il rilevamento della progressione della malattia di Alzheimer (AD). Inoltre, gli autori hanno prima applicato SHAP per selezionare le caratteristiche critiche nel classificatore. Poi, gli autori hanno utilizzato un sistema basato su regole fuzzy. SHAP è stato in grado di fornire una spiegazione locale per la previsione di diagnosi/progressione di un paziente specifico in merito all'impatto delle caratteristiche. Inoltre, il sistema basato su regole fuzzy poteva generare forme di linguaggio naturale per aiutare i pazienti e i medici a comprendere il modello AI. In [28], Peng et al. hanno proposto un framework XAI in grado di assistere i medici nella prognosi dei pazienti affetti da epatite. In questo lavoro, gli autori hanno confrontato metodi XAI intrinseci come la regressione logistica (LR), l'albero decisionale (DT) e kNN con i modelli complessi SVM, XGBoost e RF. Inoltre, gli autori hanno applicato i metodi post hoc SHAP, LIME e i diagrammi di dipendenza parziale (PDP) [29].

In [30], Sarp et al. hanno prima proposto un modello basato su CNN per la classificazione delle ferite croniche e poi hanno applicato il metodo XAI LIME per spiegare il modello basato su CNN. Il modello proposto basato su CNN ha utilizzato anche la tecnica del transfer learning e ha ottenuto una precisione media

del 95%, un richiamo medio del 94% e un punteggio medio F1 del 94%. L'immagine originale della ferita e la relativa heatmap prodotta da LIME sono mostrate nella Figura 4. Utilizzando LIME, il modello può fornire indicazioni visive al medico. Tan et al. hanno presentato una rete neurale logica per l'otosclerosi (LNN) su fette ossee temporali di tomografia computerizzata ad alta risoluzione (HRCT) per la diagnosi di otosclerosi fenestrale [31]. Il metodo proposto ha raggiunto un'AUC del 99,5% sul set di dati di prova esterno. Inoltre, hanno applicato il metodo XAI per visualizzare le rappresentazioni profonde apprese del modello LNN. In [32], Wu et al. hanno proposto un grafo multigranulare controfattuale a supporto dell'estrazione dei fatti (CMGE) per la diagnosi del linfedema. La CMGE è una rete neurale a grafo in grado di estrarre fatti dalla cartella clinica elettronica (EMR). Inoltre, può ottenere una relazione causale tra le caratteristiche. Il modello proposto è stato valutato sulle attuali cartelle cliniche elettroniche cinesi e ha dimostrato un approccio accurato e interpretabile, fornendo un ragionamento controfattuale sul grafico.



(a) **(b)**

Figura 4. Immagine della ferita cronica e relativa mappa di importanza utilizzando LIME [30]: (a) immagine originale della ferita; (b) mappa di importanza.

Analogamente, Chen et al. hanno presentato un modello di diagnosi clinica interpretabile sui papers EMR [33]. Inoltre, il modello proposto consisteva in ensemble di reti bayesiane e reti CNN entity-aware, con un'accuratezza di predizione Top-3 superiore all'88%. In particolare, la spiegabilità della rete bayesiana

nel modello è stata ottenuta costruendo connessioni tra malattie e sintomi. Successivamente, tre medici certificati hanno valutato la spiegazione del modello esaminando le relazioni estratte nel grafo della conoscenza medica. In [34], Rucco et al. hanno combinato le caratteristiche topologiche e testuali e hanno presentato un'applicazione XAI per la diagnosi del glioblastoma. Inoltre, gli autori hanno convalidato il modello di IA proposto sul recupero dell'inversione attenuata dal fluido (FLAIR) per la classificazione del glioblastoma multiforme (GBM). In termini di spiegabilità, gli autori hanno utilizzato i metodi LIME XAI per calcolare la rilevanza locale delle caratteristiche per i campioni del set di test. In [35], Gu et al. hanno proposto un sistema di diagnosi assistita dal computer chiamato VINet per fornire interpretazioni diagnostiche visive. Il VINet proposto ha ottenuto un'accuratezza di classificazione dell'82,15% su un set di dati di immagini di tomografia computerizzata (LUNA16). Inoltre, gli autori hanno confrontato VINet con altri metodi XAI come CAM, visual back-propagation (VBP) e layer-wise relevance propagation (LRP). Inoltre, è stato in grado di dimostrare le interpretazioni visive SOTA.

In [36], Kroll et al. hanno sviluppato un framework basato sull'evoluzione grammaticale per la diagnosi e la prognosi della malattia di Alzheimer (AD). Il framework proposto è stato valutato su un set di dati di risonanza magnetica (MRI) e i risultati sperimentali hanno dimostrato che è in grado di fornire sia accuratezza che spiegabilità. L'evoluzione grammaticale è riuscita a generare modelli di stringhe in base alle regole di produzione. In termini di spiegabilità, gli autori hanno utilizzato l'evoluzione grammaticale per la rappresentazione delle caratteristiche e le hanno poi applicate alla classificazione. Meldo et al. hanno proposto un sistema di diagnosi computerizzata del cancro ai polmoni con frasi esplicative [37]. Il sistema proposto è composto da due parti: la prima è un modello XAI locale post hoc che utilizza LIME; la seconda trasforma le caratteristiche importanti selezionate in linguaggio naturale. In [38], Yeboah et al. hanno presentato un modello XAI basato sul clustering ensemble per

l'analisi prognostica e diagnostica delle lesioni cerebrali traumatiche (TBI). Inoltre, il framework spiegabile proposto è in grado di combinare l'analisi automatizzata dei dati e le conoscenze mediche degli esperti. Per quanto riguarda l'interpretazione, il framework ha utilizzato la valutazione della qualità delle caratteristiche di clustering, l'identificazione delle caratteristiche discriminanti e l'interpretazione clinica. In [39], Wang et al. hanno proposto un modello basato su CNN denominato COVID-Net per il rilevamento di casi di COVID-19 utilizzando immagini radiografiche del torace (CXR). Gli autori hanno anche confrontato COVID-Net con VGG-19 e ResNet-50. COVID-Net ha ottenuto un'accuratezza del 93,3% sul dataset di test COVIDx e una sensibilità del 91%. Inoltre, gli autori hanno applicato il metodo XAI GSInquire [40] per studiare la predizione di COVID-Net. GSInquire può essere utilizzato per ottenere una migliore comprensione delle reti neurali e può anche imparare a generare reti.

In [41], Sabol et al. hanno proposto un'applicazione XAI denominata cumulative fuzzy class membership criterion (CFCMC) che potrebbe assistere i patologi ed essere utilizzata per la diagnosi del cancro colorettale. Il sistema proposto fornisce una spiegazione semantica dei risultati della classificazione del tipo di tessuto. Inoltre, CFCMC ha mostrato le immagini originali dell'intero vetrino (WSI) del tessuto e la corrispondente visualizzazione della mappa delle etichette. Il modello XAI proposto è stato valutato da 14 patologi. Wei et al. hanno progettato una diagnosi assistita dall'intelligenza artificiale dei noduli tiroidei e poi hanno testato il modello per le prestazioni di classificazione [42]. Inoltre, gli autori hanno applicato tecniche di pre-elaborazione dei dati per localizzare e diagnosticare i noduli tiroidei. Attraverso esperimenti, hanno scoperto che il rapporto A/T e le informazioni sul margine dei noduli tiroidei sono caratteristiche cliniche importanti per la diagnosi dei noduli tiroidei. Gli autori hanno utilizzato la mappatura dell'attivazione di classe per visualizzare la rete neurale basata su CNN proposta per spiegare il modello. Il CAM utilizza il pooling della media globale e uno strato completamente connesso per visualizzare la

rete neurale e le caratteristiche più importanti. In [43], Chang et al. hanno presentato una rete neurale profonda spiegabile (EDNN). Il modello è stato addestrato su un set di dati con 200 pazienti schizofrenici e controlli sani della coorte Taiwan Aging and Mental Illness (TAMI). Utilizzando la coorte TAMI, il framework proposto ha raggiunto un'accuratezza dell'84,0% nella materia grigia (GM) e del 90,22% nella materia bianca (WM). In termini di spiegabilità, il sistema ha fornito una visualizzazione tridimensionale (3D) dei dati di imaging cerebrale del soggetto, in grado di ottimizzare il processo diagnostico. In [44], Magesh et al. hanno proposto un modello basato su CNN per il rilevamento precoce della malattia di Parkinson (PD). Il dataset del lavoro consisteva in 642 immagini di tomografia computerizzata a emissione di fotoni singoli (SPECT) provenienti dal database Parkinson's Progression Markers Initiative (PPMI). Inoltre, gli autori hanno utilizzato il transfer learning sul modello basato su CNN. Per l'interpretazione della ricerca è stato utilizzato il metodo XAI post hoc LIME. LIME ha potuto evidenziare le regioni di interesse nell'immagine SPECT con le aree di impatto che classificano i dati come controlli sani.

In [45], Cho et al. hanno proposto un metodo di apprendimento automatico interpretabile per prevedere le disposizioni di dimissione ospedaliera post-ictus. Gli autori hanno scelto il modello lineare di regressione logistica (LR) come modello di base e lo hanno poi confrontato con il modello black-box, comprendente RF, RF con AdaBoost e MLP. Per interpretare il modello black-box, l'autore ha utilizzato LIME e ha fornito spiegazioni per la previsione. Utilizzando LIME, gli autori hanno identificato le caratteristiche più importanti per il modello. In particolare, caratteristiche come l'età, il diabete e la fonte di ricovero erano importanti per la previsione delle disposizioni di dimissione ospedaliera post-ictus. Lamy et al. hanno presentato un approccio di ragionamento visivo basato sui casi (CBR) per la diagnosi del cancro al seno [46]. Il CBR utilizza un database di casi risolti in precedenza per determinare la risposta a un caso di domanda. Si tratta di una forma di ragionamento analogico. Dal

database sono stati recuperati casi simili e le loro soluzioni sono state adattate alla domanda. L'algoritmo proposto ispirato alle scatole arcobaleno automatiche (RIBA) è stato confrontato con kNN e kNN ponderato in base alla distanza (WkNN). Inoltre, il metodo proposto ha ottenuto un'accuratezza dell'80,3% su un set di dati reali sul cancro al seno. Inoltre, esperti medici hanno valutato l'applicazione XAI proposta e hanno trovato l'approccio CBR visivo molto interessante. In [47], Das et al. hanno proposto un modello XAI per la diagnosi della malattia di Alzheimer (AD), denominato sparse high-order interaction model with rejection option (SHIMR). Il modello SHIMR proposto è stato convalidato su un set di dati di AD e ha dimostrato di avere un'elevata accuratezza, interpretabilità ed economicità. Utilizzando SHIMR, è stato possibile creare set decisionali altamente accurati e interpretabili, con raccolte di regole "se-allora" [IF-THEN] che riflettono le interazioni di ordine superiore tra un insieme di caratteristiche individuali.

3.2 INTERVENTI CHIRURGICI

In [48], Yoo et al. hanno presentato un modello multiclasse XGBoost per selezionare l'opzione di chirurgia laser a livello esperto. Gli autori hanno convalidato il metodo proposto su soggetti sottoposti a chirurgia refrattiva presso il B&VIIT Eye Center e hanno ottenuto un'accuratezza del 78,9% sul set di dati di validazione esterna. Questo fornisce anche una comprensione clinica del metodo di deep learning che utilizza la tecnica SHAP. In [49], Mirchi et al. hanno proposto un framework che può essere utilizzato per l'addestramento chirurgico con feedback visivo didattico automatizzato. Gli autori hanno addestrato e valutato un modello SVM su dati medici e chirurgici simulati, ottenendo un'accuratezza del 92%, una specificità dell'82% e una sensibilità del 100%. Inoltre, hanno fornito una spiegazione approfondita dell'algoritmo di apprendimento automatico proposto, identificando le metriche insegnabili che contribuiscono alla classificazione. Fawaz et al. hanno presentato un'applicazio-

ne medica accurata e interpretabile per la valutazione delle abilità chirurgiche, addestrando una rete neurale completamente convoluzionale (FCN) per classificare i livelli di abilità chirurgica utilizzando la cinematica chirurgica [50]. Il modello proposto ha ottenuto prestazioni all'avanguardia sul dataset pubblico JIGSAWS per tre compiti di abilità chirurgica. Inoltre, gli autori hanno applicato il CAM per fornire un feedback di classificazione interpretabile. Il feedback visivo fornito dal CAM è illustrato nella Figura 5 qui sotto. Il CAM è un metodo di spiegazione visiva post hoc XAI utilizzato per i modelli basati su CNN per individuare le caratteristiche della CNN che influenzano le decisioni di classificazione. In particolare, il CAM utilizza uno strato di pooling medio globale (GAP) dopo lo strato convoluzionale, con la possibilità di visualizzare quali parti del processo contribuiscono maggiormente alla classificazione delle abilità. Attraverso l'indagine del comportamento specifico di un livello di abilità, gli osservatori possono identificare i modelli di movimento caratteristici di una particolare classe di chirurghi.

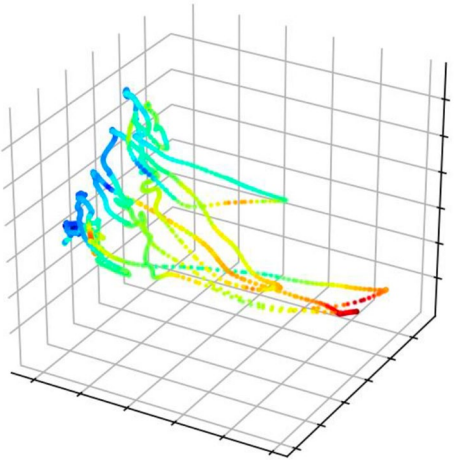


Figura 5. Feedback visivo del compito chirurgico del chirurgo con CAM [50]. Feedback visivo del compito chirurgico del chirurgo con CAM [50]. Le sottosequenze rosse e arancioni nel grafico mostrano l'elevato contributo al compito di valutazione dell'abilità chirurgica del chirurgo. Al contrario, le sottosequenze verdi e blu indicano il basso contributo.

In [51], Kletz et al. hanno proposto un'applicazione medica basata su CNN in grado di apprendere la rappresentazione degli strumenti in laparoscopia e hanno convalidato il modello su vari set di dati. Hanno anche fornito mappe di attivazione dei diversi strati della CNN per aiutare a capire come il modello ha classificato lo strumento. Un sistema di intelligenza artificiale spiegabile, XAI-CBIR, è stato proposto da Chittajallu et al. per la formazione chirurgica [52]. XAI-CBIR è un metodo XAI post hoc di spiegazione tramite esempi. Fornisce spiegazioni estraendo gli esempi rappresentativi. In particolare, sfrutta un modello di deep learning auto-supervisionato per estrarre descrittori semantici dai fotogrammi video MIS. Inoltre, utilizza una mappa di salienza per fornire una spiegazione visiva del motivo per cui il sistema ritiene che l'immagine recuperata sia simile all'immagine richiesta. Utilizzando il sistema XAI-CBIR, i video di chirurgia mini-invasiva (MIS) possono essere recuperati in base al loro contenuto. Il sistema proposto è stato valutato sul dataset Cholec80 e la percentuale di fotogrammi rilevanti tra i primi 50 fotogrammi recuperati per tre fasi è stata rispettivamente del 64,42%, 99,54% e 99,09%. Inoltre, è stata applicata una mappa di salienza per guidare il feedback di rilevanza con spiegazioni visive.

4. DISCUSSIONE

4.1 TENDENZE ATTUALI DELLA RICERCA

Le tecniche di intelligenza artificiale, come il deep learning , hanno recentemente svolto un ruolo rivoluzionario nell'assistenza sanitaria, compresa la diagnosi e la chirurgia. Queste tecniche hanno dimostrato di essere efficaci in questi campi. L'accuratezza di alcuni compiti di diagnosi basati sul deep learning supera persino gli esperti medici umani. Tuttavia, la natura black-box del modello di deep learning limita la spiegabilità di questi modelli e ne limita l'impiego pratico in medicina. Nel campo interdisciplinare dell'intelligenza artificiale e della medicina, molti ricercato-

ri hanno capito che la chiave dell'impiego dell'IA nell'ambiente clinico non è l'accuratezza del modello, ma la sua spiegabilità. Le applicazioni di IA in campo medico devono essere spiegate prima di essere accettate e integrate nella pratica medica. Di conseguenza, l'accettazione delle applicazioni di IA in campo medico richiede un'intelligenza artificiale spiegabile e ciò motiva l'indagine sulla XAI in campo medico.

In questa indagine sono stati inclusi 27 lavori sulla diagnosi e sulla chirurgia che utilizzano l'intelligenza artificiale spiegabile. Come osservato nella Tabella 1 e nella Tabella 2, gli studi inclusi in questa indagine sono stati analizzati dal punto di vista dell'algoritmo di IA, del metodo XAI e delle prestazioni dell'IA. Abbiamo scoperto che l'algoritmo di apprendimento automatico tradizionale più popolare è la foresta casuale, con il 25,9% (7/27) degli articoli pubblicati riportati in questa indagine che hanno condotto esperimenti con la foresta casuale; il metodo di apprendimento profondo più popolare è quello delle reti neurali convoluzionali (CNN), con il 37,0% (10/27) degli articoli che utilizzano un modello basato su CNN come VGG-16; inoltre, LIME è l'approccio XAI più comunemente utilizzato in questi articoli, con il 25,9% (7/27) degli articoli che utilizzano LIME per spiegare il modello di apprendimento automatico proposto. Per quanto riguarda i metodi XAI, la maggior parte degli articoli ha utilizzato metodi post hoc e ha seguito la pipeline introdotta nella Figura 3. In primo luogo, sono stati applicati algoritmi di deep learning per spiegare il modello di Machine Learning proposto. In primo luogo hanno applicato algoritmi di deep learning come modelli basati su CNN o modelli complessi di machine learning random forest e poi hanno utilizzato metodi post hoc come LIME, SHAP e PDP per spiegare il modello di IA; in secondo luogo, hanno costruito le applicazioni mediche e fornito il processo decisionale ai medici. Inoltre, per quanto riguarda la valutazione XAI, solo l'11,1% (3/27) degli studi ha applicato la valutazione XAI.

SN#	Riferimento	Anno	Obiettivo	Algoritmo AI	Metriche di valutazione dell'IA	Metodo XAI	Metodo XAI Tipo	Valutazione XAI?
1	[24]	2021	Diagnosi di allergia	kNN, SVM, C 5.0, MLP, AdaBag, RF	Accuratezza: 86,39%. Sensibilità: 75%	Regole di previsione delle condizioni (IF-THEN)	Basato su regole	No
2	[25]	2021	Terapie per il cancro al seno	Analisi dei cluster	N/D	Adattivo riduzione delle dimensioni	Riduzione delle dimensioni	No
3	[26]	2021	Dorsale	SVM a una classe, RF binario	F1: 80 ± 12% MCC: 57 ± 23% BSS: 33 ± 28%	Spiegazioni locali interpretabili modello-agnostiche (LIME)	Spiegazione per semplificazione	No
4	[27]	2021	Malattia di Alzheimer	Modello a due strati con RF	Primo strato: accuratezza: 93,95%. Punteggio F1: 93,94%. Secondo strato: 87.08% Punteggio F1: 87,09%.	SHAP, Fuzzy	Rilevanza delle caratteristiche, basata su regole	No
5	[28]	2021	Epatite	LR, DT, kNN, SVM, RF	Accuratezza: 91,9%.	SHAP, LIME, diagrammi di dipendenza parziale (PDP)	Rilevanza delle caratteristiche, spiegazione per semplificazione	No
6	[30]	2021	Ferita cronica	Modello basato su CNN: preaddestrato VGG-16	Precisione: 95% Richiamo: 94% Punteggio F1: 94%	LIME	Spiegazione per semplificazione	No

7	[31]	2021	Otosclerosi fenestrale	Modello basato su CNN: proposto modello di rete neurale logica (LNN) per l'otosclerosi	AUC: 99.5% Sensibilità: 96,4%. Specificità: 98,9%.	Visualizzazione delle rappresentazioni profonde apprese	Spiegazione visiva	No
8	[32]	2021	Linfedema (EMR cinese)	Metodo di estrazione dei fatti a supporto del grafo multigranulare controfattuale (CMGE)	Precisione: 99,04%. Richiamo: 99.00% Punteggio F1: 99,02%.	Rete neurale grafica, ragionamento controfattuale	Architettura della rete neurale ristretta	No
9	[33]	2020	Diagnosi clinica	Reti neurali convoluzionali consapevoli delle entità (ECNN)	Sensibilità della Top-3: 88,8%.	Insiemi di reti bayesiane	Modelli bayesiani	Si
10	[34]	2020	Diagnosi di glioblastoma multiforme (GBM)	VGG16	Accuratezza: 97%.	LIME	Spiegazione per semplificazione	No
11	[35]	2020	Diagnostica dei noduli polmonari	CNN	Accuratezza: 82,15%.	Rete interpretabile visivamente (VINet), LRP, CAM, VBP	Spiegazione visiva	No
12	[36]	2020	Diagnosi della malattia di Alzheimer	Naïve Bayes (NB), evoluzione grammaticale	ROC: 0,913 Precisione: 81,5%. Punteggio F1: 85,9% Bacca: 0,178	Grammatica libera dal contesto (CFG)	Basato su regole	No
13	[37]	2020	Diagnosi di cancro al polmone	Reti neurali, RF	N/D	LIME, spiegazione in linguaggio naturale	Spiegazione per semplificazione, spiegazione del testo	No

14	[38]	2020	Identificazione delle lesioni cerebrali traumatiche (TBI)	k-means, clustering spettrale, miscela gaussiana	N/D	Valutazione della qualità delle caratteristiche di clustering	Rilevanza delle caratteristiche	No
15	[39]	2020	COVID-19 diagnosi radiografica del torace	Modello basato su CNN: rete COVID proposta	Precisione: 93,3%. Sensibilità: 91,0%.	GSInquire	Architettura della rete neurale ristretta	No
16	[41]	2020	Diagnosi di cancro del colon-retto	CNN	Precisione: 91,08%. Precisione: 91,44%. Richiamo: 91.04% Punteggio F1: 91,26%.	Criterio di appartenenza alla classe fuzzy cumulativa spiegabile (X-CFCMC)	Spiegazione visiva	Sì
17	[42]	2020	Diagnosi dei noduli tiroidei	Rete neurale	Accuratezza: 93,15%. Sensibilità: 92,29%. Specificità: 93,62%.	CAM	Spiegazione visiva	No
18	[43]	2020	Fenotipizzazione della diagnosi dei disturbi psichiatrici	DNN	Accuratezza della materia bianca: 90,22%. Sensibilità: 89,21%. Specificità: 91,23%.	Rete neurale profonda spiegabile (EDNN)	Spiegazione visiva	No
19	[44]	2020	Diagnosi della malattia di Parkinson (PD)	CNN	Accuratezza: 95,2%. Sensibilità: 97,5%. Specificità: 90,9%.	LIME	Spiegazione per semplificazione	No

20	[45]	2019	Dimissione ospedaliera post-ictus disposizione	LR, RF, RF con AdaBoost, MLP	Accuratezza del test: 71%. Precisione: 64% Richiamo: 26% Punteggio F1: 59%	LR, LIME	Intrinseco, Spiegazione per semplificazione	No
21	[46]	2019	Decisione diagnostica e decisione terapeutica sul cancro al seno	kNN, kNN ponderato sulla distanza (WkNN), algoritmo ispirato alle scatole arcobaleno (RBIA)	Precisione: 80,3%.	Approccio al ragionamento basato sui casi (CBR)	Spiegazione con un esempio	Si
22	[47]	2019	Diagnosi di Alzheimer	RF, SVM, DT	Sensibilità: 84%. Specificità: 67%. AUC: 0.81	Un modello ML interpretabile: sparse high-order interaction model with rejection option (SHIMR)	Basato su regole	No

Tabella 1. Rassegna della letteratura sulle applicazioni XAI in campo medico per la diagnosi. SN#: numero di serie; N/A: non applicabile; AI: intelligenza artificiale; XAI: intelligenza artificiale spiegabile; kNN: k-nearest neighbor; SVM: support vector machine; MLP: multi-layer perceptron; RF: random forest; MCC: matthews correlation coefficient; BSS: brier skill score; SHAP: SHapley Additive exPlanations; LR: regressione logistica; DT: albero decisionale; LIME: Local interpretable model-agnostic explanations; PDP: partial dependence plots; CNN: convolutional neural networks; DNN: deep neural network; AUC: area sotto la curva.

SN#	Riferimento	Anno	Obiettivo	Algoritmo AI	Metriche di valutazione dell'IA	Metodo XAI	Metodo XAI Tipo	Valutazione XAI?
23	[48]	2020	Raccomandazioni basate sull'evidenza per la chirurgia	XGBoost	Accuratezza della convalida: 78,9%.	SHAP	Rilevanza delle caratteristiche	No

24	[49]	2020	Formazione in chirurgia	SVM	Accuratezza: 92%. Sensibilità: 100% Specificità: 82%.	Assistente operativo virtuale	Rilevanza delle caratteristiche	No
25	[50]	2019	Valutazione dell'abilità chirurgica	FCN	Precisione della sutura: 100%. Precisione del passaggio dell'ago: 100%. Precisione di legatura dei nodi: 92,1%.	CAM	Spiegazione visiva	No
26	[51]	2019	Riconoscimento automatico degli strumenti nei video di laparoscopia	CNN	M2CAI Sintonizzazione dei dati Cholec su InstCnt non strumento Strumento: Precisione: 96%. Sensibilità: 86%. Punteggio F1: 97%.	Mappe di attivazione	Spiegazione visiva	No
27	[52]	2019	Formazione chirurgica	CNN	Percentuale di fotogrammi rilevanti tra i primi 50 fotogrammi recuperati per tre fasi: 64.42%, 99.54%, 99.09%	Mappa di salienza, recupero di immagini basato sul contenuto	Spiegazione visiva, spiegazione tramite esempi	No

Tabella 2. Rassegna della letteratura sulle applicazioni XAI in campo medico e chirurgico. SN#: numero di serie; AI: intelligenza artificiale; XAI: intelligenza artificiale spiegabile; SHAP: SHapley Additive exPlanations; SVM: support vector machine; FCN: fully convolutional neural network; CAM: class activation mapping; CNN: convolutional neural networks.

Il riassunto di questa indagine suggerisce che diversi metodi di apprendimento automatico o di deep learning sarebbero soluzioni ottimali per varie applicazioni XAI in campo medico. Non esiste un modello di apprendimento automatico o un approccio XAI unificato che si adatti a tutte le attività di diagnosi e chirurgia e dipende dalle dimensioni del set di dati, dal tipo di dati e da molti altri fattori.

4.2 VETRINA SPERIMENTALE: DIAGNOSI DEL CANCRO AL SENO

Abbiamo analizzato le tendenze attuali della ricerca riassumendo la letteratura inclusa nell'indagine. Inoltre, per comprendere meglio i metodi XAI, abbiamo mostrato una vetrina sperimentale di un'applicazione XAI medica, ovvero la diagnosi del cancro al seno.

4.2.1 Set di dati

Una delle forme di cancro più comuni tra le donne è il cancro al seno. In questo lavoro, utilizziamo un dataset Wisconsin (Diagnostic) contenente 569 pazienti [52]. Il dataset è disponibile pubblicamente presso il repository di machine learning dell'UCI: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 (accesso il 22 dicembre 2021). Le caratteristiche del dataset sono numeriche ed estratte dall'immagine digitalizzata dell'aspirato con ago sottile (FNA) di una massa mammaria. In termini di distribuzione delle classi, ci sono 357 benigne e 212 maligne.

4.2.2 Impostazione dell'esperimento

Gli esperimenti sono stati eseguiti su un computer portatile con una CPU a 6 core da 2,6 GHz e sono stati implementati in Python: Il 70% del dataset del cancro al seno è stato utilizzato per l'addestramento e il 30% del dataset è stato utilizzato per i test. Il modello black-box proposto è stato addestrato utilizzando il toolkit Scikit-learn [53]. Per i metodi XAI, abbiamo utilizzato la libreria Python InterpretML [54].

4.2.3 Metodo XAI intrinseco: Basato su regole

I metodi basati su regole sono espressioni logiche della forma IF-THEN. In questa vetrina abbiamo implementato una regola basata sul compito di classificazione del cancro al seno. Inoltre, abbiamo valutato le prestazioni del modello in termini di accuratezza, precisione, richiamo e punteggio F1. Il modello basato

su regole proposto ha ottenuto un'accuratezza del 60,81%, una precisione del 60,95%, un richiamo del 99,04% e un punteggio F1 del 75,46%.

4.2.4 Metodo XAI post hoc: SHAP

Il tipo di dataset era numerico e abbiamo scelto una foresta casuale con 300 alberi come modello black-box. La foresta casuale ha ottenuto un'accuratezza del 95,91%, una precisione del 97,09%, un richiamo del 96,15% e un punteggio F1 del 96,62%. La foresta casuale ha ottenuto risultati migliori rispetto al metodo basato sulle regole. Inoltre, abbiamo applicato il metodo post hoc SHAP per spiegare il modello black-box. SHAP è un metodo XAI post hoc sulla rilevanza delle caratteristiche, ispirato alla teoria dei giochi. Mira ad aumentare l'interpretabilità calcolando il valore di importanza di ogni caratteristica per ogni previsione utilizzando i valori di Shapley. La Figura 6 mostra come SHAP interpreta le previsioni del modello black-box.

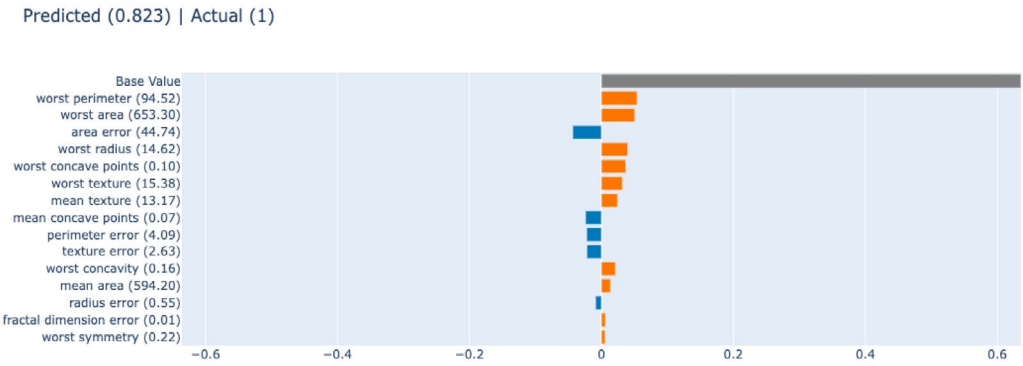


Figura 6. Interpretazione di una previsione con il metodo XAI post hoc: SHAP.

4.2.5 Metodo XAI post hoc: LIME

LIME è un metodo XAI di spiegazione per semplificazione post hoc, che può spiegare una singola previsione generata da

qualsiasi modello black-box. Spiega una previsione sostituendo il modello complesso con un modello surrogato interpretabile localmente. Concentrandosi su una superficie decisionale sufficientemente ristretta, LIME cerca di modellare esplicitamente il vicinato locale di qualsiasi previsione. La Figura 7 è un esempio di LIME per interpretare la previsione di un modello black-box. In particolare, le caratteristiche "area peggiore", "raggio peggiore", "perimetro peggiore", "struttura peggiore", "concavità peggiore", "struttura media" e "area media" hanno un effetto positivo sulla predizione.

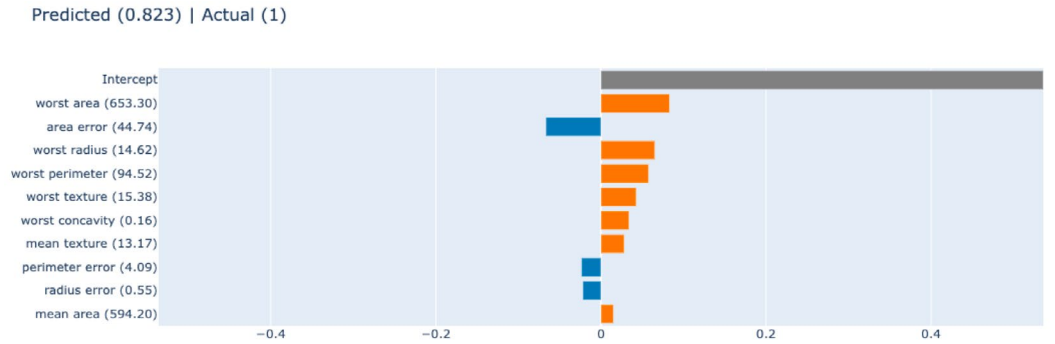


Figura 7. Interpretazione di una previsione con il metodo XAI post hoc: LIME. L'asse delle ascisse mostra l'effetto delle caratteristiche.

4.2.6 Metodo XAI post hoc: PDP

Il PDP è anche un metodo XAI post hoc di spiegazione della rilevanza delle caratteristiche, che interpreta il modello black-box tracciando l'impatto di sottoinsiemi di caratteristiche sulle previsioni del modello. La Figura 8 è una visualizzazione PDP della dipendenza tra la caratteristica "raggio medio" e la risposta.

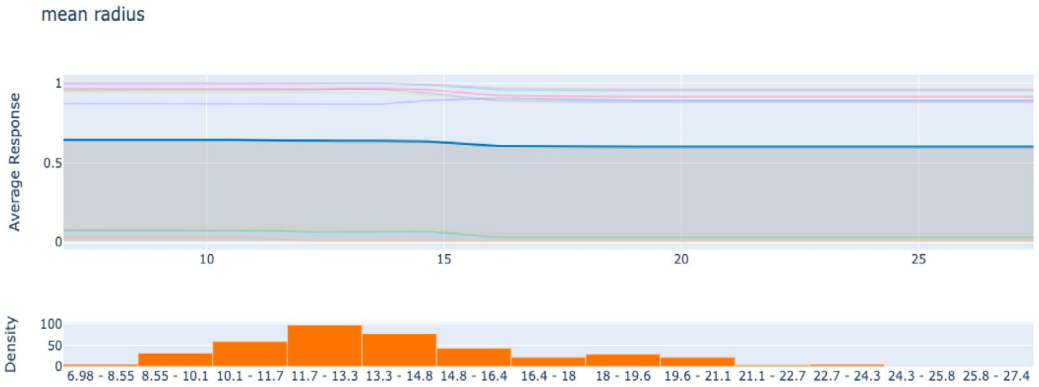


Figura 8. Interpretazione delle decisioni del modello black-box con PDP per la caratteristica "raggio medio".

4.3 SFIDE, LIMITI E LACUNE DELLA RICERCA

L'indagine ha inoltre identificato alcune sfide e limitazioni. In primo luogo, in alcuni papers l'accuratezza è stata l'unica metrica di valutazione dell'apprendimento automatico utilizzata per valutare le prestazioni del modello, il che è irragionevole. L'uso di una sola metrica di valutazione non può fornire una valutazione oggettiva dell'algoritmo di apprendimento automatico. In secondo luogo, la dimensione dei dataset nelle applicazioni XAI mediche era relativamente piccola e la qualità dei dati non era garantita. Di conseguenza, le prestazioni del modello di intelligenza artificiale possono essere limitate dalle dimensioni ridotte dei dataset e dalla bassa qualità dei dati di input. Il modello di intelligenza artificiale è stato addestrato e convalidato solo su un set di dati di dimensioni ridotte; di conseguenza, è probabile che il modello abbia causato un problema di overfitting. La generalizzazione del modello era bassa. In terzo luogo, per quanto riguarda la valutazione XAI, non esistono ancora metodi unificati di valutazione XAI accettati dalla maggior parte dei ricercatori della comunità. I modelli XAI possono essere valutati solo qualitativamente perché la valutazione si basa ancora sulla

cognizione umana. Tuttavia, la maggior parte dei lavori di questa indagine ha fornito solo i metodi XAI senza alcuna valutazione XAI. Solo pochi ricercatori hanno fornito valutazioni XAI da parte di medici professionisti. Infine, molti studi hanno applicato solo i metodi di apprendimento automatico o XAI esistenti. Questi approcci di IA sono stati progettati senza la partecipazione di esperti medici, il che ha fatto sì che queste applicazioni XAI mediche mancassero di innovazione e di conoscenze preliminari da parte dei medici e potessero non soddisfare le reali esigenze cliniche dei medici.

Abbiamo inoltre concluso con due lacune nella ricerca che abbiamo riscontrato dopo aver esaminato la letteratura sull'XAI medica. In primo luogo, la maggior parte degli studi nel campo interdisciplinare dell'intelligenza artificiale e della medicina si concentra sui metodi di deep learning, tra cui MLP, CNN, RNN e trasformatori. Questi modelli basati sul deep learning, come i trasformatori, contengono milioni di parametri e sono difficili da interpretare. Tuttavia, gli scienziati dei dati e gli esperti di IA nel campo interdisciplinare dell'intelligenza artificiale e della medicina dovrebbero concentrarsi su XAI piuttosto che sui modelli di deep learning SOTA. In secondo luogo, le applicazioni XAI in campo medico dovrebbero essere valutate da esperti del settore. Tuttavia, la maggior parte delle applicazioni XAI in campo medico manca di valutazione XAI e di valutazione da parte di esperti medici. Le applicazioni XAI mediche dovrebbero avere un'interfaccia HCI ben progettata e fornire una spiegazione ragionevole agli esperti medici.

4.4 DIREZIONI FUTURE

Per quanto riguarda le direzioni future della ricerca, riteniamo che l'IA sarà applicata in molte diagnosi e attività chirurgiche diverse. L'IA svolgerà un ruolo essenziale perché può aumentare la trasparenza di questi modelli e ottenere la fiducia dei medici. Per affrontare le sfide di cui sopra, inizialmente riteniamo che sarebbe auspicabile considerare diverse metriche di valutazione,

come la specificità e la sensibilità, per valutare il deep learning, e non solo l'accuratezza. Inoltre, sarebbe preferibile utilizzare la convalida incrociata per validare il modello addestrato. Inoltre, in futuro, dovremmo raccogliere e costruire il set di dati da più fonti, ad esempio da diversi ospedali, per aumentare le dimensioni del set di dati e migliorare la capacità di generalizzazione del modello di apprendimento automatico.

Inoltre, possiamo anche applicare l'apprendimento automatico federato per mantenere al sicuro i dati medici identificabili. Inoltre, alcune tecniche come l'aumento dei dati [data augmentation] [54], l'apprendimento per trasferimento [transfer learning] [55] e l'apprendimento a pochi scatti [few-shot learning] [56] potrebbero essere prese in considerazione per affrontare il problema delle dimensioni ridotte dei dataset. In terzo luogo, in termini di valutazione XAI, non esiste un vero e proprio consenso. Non sono state adottate metriche di valutazione oggettive e unificate. Per le valutazioni XAI generali, alcuni ricercatori hanno proposto un approccio di valutazione. Ad esempio, Holzinger et al. hanno presentato un nuovo approccio per spiegare la qualità, chiamato scala di causalità del sistema (SCS) [53]. Utilizzava il metodo della scala Likert e poteva valutare rapidamente se il modello spiegabile era appropriato per lo scopo previsto. Tuttavia, per la valutazione delle XAI mediche, riteniamo che questa debba basarsi su una valutazione centrata sull'uomo. Più precisamente, dovrebbe essere valutata sia da esperti di medicina sia da esperti di IA. Ad esempio, gli esperti di medicina possono valutare il metodo XAI utilizzando i compiti clinici correlati per garantire che l'applicazione XAI medica possa fare un'inferenza clinica spiegabile. In confronto, gli esperti di IA possono valutare le applicazioni XAI in base alla loro generalizzazione e robustezza.

Infine, i professionisti del settore medico dovrebbero partecipare alle fasi di progettazione e sviluppo dei futuri studi sulle applicazioni XAI in campo medico. Una buona applicazione XAI medica richiede una collaborazione interdisciplinare. In particolare, gli esperti di medicina dovrebbero fornire le loro conoscenze mediche pregresse e i loro suggerimenti e feedback

contribuiranno a migliorare la progettazione degli algoritmi di IA. Gli esperti di IA e i data scientist devono assicurarsi che le applicazioni XAI mediche possano aiutare gli esperti medici a fare inferenze cliniche spiegabili. Di conseguenza, riteniamo che i modelli XAI medici diventeranno più accettabili per il settore medico. Per raggiungere questo obiettivo, il miglioramento delle interazioni uomo-computer (HCI) è un approccio promettente. Sarà possibile per gli esperti di medicina e di IA lavorare insieme attraverso un'applicazione medica HCI ben progettata.

4.5 DOMANDE DI RICERCA

Le domande di ricerca poste nell'introduzione di questa revisione sono discusse qui.

- RQ1: Quali sono le attuali tendenze della ricerca sulle applicazioni XAI in campo medico?
- Risposta: Sulla base della letteratura inclusa in questa revisione, abbiamo riscontrato che la maggior parte degli studi in letteratura ha applicato metodi XAI post hoc. In generale, hanno seguito la pipeline che abbiamo illustrato nella Figura 3.
- RQ2: In che modo gli studi inclusi in questa revisione affrontano il trade-off tra accuratezza e spiegabilità?
- Risposta: Abbiamo riassunto gli studi analizzati e abbiamo elencato le metriche di valutazione dell'IA e le valutazioni XAI. In termini di prestazioni dell'IA, la maggior parte degli studi ha ottenuto buoni risultati. Tuttavia, solo pochi studi hanno fornito valutazioni XAI. Inoltre, la maggior parte dei lavori non ha valutato l'efficacia del modello da parte di esperti medici. Pertanto, non possiamo rispondere a come questi studi affrontano il trade-off tra accuratezza e spiegabilità.
- RQ3: È possibile implementare questi modelli in un ambiente clinico reale per assistere gli esperti medici nel fare inferenze cliniche spiegabili?

- Risposta: Attualmente, le applicazioni XAI in campo medico presentano ancora molte limitazioni e non è possibile implementare i modelli nell'ambiente clinico. Tuttavia, riteniamo che la direzione futura delle applicazioni mediche XAI sia promettente.

5. CONCLUSIONI

In conclusione, questo articolo ha esaminato la letteratura esistente e ha fornito un'indagine approfondita sulle applicazioni XAI in campo medico nella diagnosi e nella chirurgia. I metodi di IA, i metodi XAI, il tipo di XAI, la valutazione XAI e le prestazioni di IA degli articoli inclusi nell'indagine sono stati discussi e confrontati. Inoltre, abbiamo presentato una vetrina sperimentale per illustrare come diversi metodi XAI possano essere utilizzati in applicazioni XAI in campo medico. Inoltre, abbiamo fornito una sintesi dello studio e abbiamo affrontato i limiti attuali e le prospettive future delle applicazioni XAI in campo medico. Nel campo interdisciplinare dell'intelligenza artificiale e della medicina, riteniamo che questa rassegna possa ridurre il divario tra professionisti dell'IA e medici e fornire informazioni utili ai futuri ricercatori per la progettazione di applicazioni XAI in campo medico.

8 - INTELLIGENZA ARTIFICIALE IN ONCOLOGIA: APPLICAZIONI ATTUALI E PROSPETTIVE FUTURE

Tratto e tradotto da
Luchini, C., Pea, A. & Scarpa, *Artificial intelligence in oncology: current applications and future perspectives, Br J Cancer* 126, 4-9 (2022).

https://doi.org/10.1038/s41416-021-01633-1

Le parti omesse dal curatore rispetto all'originale sono indicate dal segno **[...]**

ABSTRACT

L'intelligenza artificiale (AI) sta concretamente ridisegnando il panorama e gli orizzonti dell'oncologia, aprendo nuove importanti opportunità per migliorare la gestione dei pazienti oncologici. Analizzando i dispositivi basati sull'IA che hanno già ottenuto l'approvazione ufficiale da parte della Federal Drug Administration (FDA), dimostriamo che la diagnostica del cancro è l'area oncologica in cui l'IA è già entrata con il maggiore impatto nella pratica clinica. Inoltre, i tumori al seno, ai polmoni e alla prostata rappresentano i tipi di cancro specifici che ora stanno sperimentando maggiori vantaggi dai dispositivi basati sull'IA. Le prospettive future dell'IA in oncologia sono discusse: le sfide più importanti per portare a termine la "rivoluzione dell'IA" in oncologia sono la creazione di piattaforme multidisciplinari, la comprensione dell'importanza di tutte le neoplasie, compresi i tumori rari, e il continuo supporto per garantire la crescita dell'IA.

1. INTRODUZIONE

L'intelligenza artificiale (IA) sta concretamente rimodellando le nostre vite ed è giunto il momento di comprenderne l'evoluzione e i risultati per modellare le strategie di sviluppo future. Questo vale anche per l'oncologia e i campi correlati, dove l'IA sta aprendo nuove importanti opportunità per migliorare la gestione dei pazienti oncologici, come verrà evidenziato in questo documento di prospettiva.

Nel 1950, Alan Turing fu il primo a concepire l'idea di utilizzare i computer per imitare il comportamento intelligente e il pensiero critico [1]. Nel 1956, John McCarthy coniò il termine "intelligenza artificiale" come "scienza e ingegneria della creazione di macchine intelligenti" [1, 2]. L'IA, nata come una semplice serie di regole "se, allora", è progredita negli anni successivi fino a comprendere algoritmi variegati e compositi che funzionano in modo simile al cervello umano [1].

Oggi l'IA rappresenta un modello emergente e in rapida evoluzione che riguarda diversi ambiti scientifici, anche quelli dedicati alla gestione dei pazienti oncologici [2,3,4,5]. Può essere vista come un concetto generale che indica la capacità di una macchina di apprendere e riconoscere modelli e interazioni da un numero sufficiente di modelli rappresentativi, e di utilizzare queste informazioni per migliorare l'approccio attuale al processo decisionale in un campo specifico [3,4,5].

Nell'oncologia di precisione, l'IA sta ridisegnando lo scenario esistente, con l'obiettivo di integrare la grande quantità di dati derivati da analisi multiomiche con gli attuali progressi nel calcolo ad alte prestazioni e con strategie innovative di deep-learning [3]. In particolare, le applicazioni dell'IA si stanno espandendo e includono nuovi approcci per la rilevazione, lo screening, la diagnosi e la classificazione del cancro, la caratterizzazione della genomica del cancro, l'analisi del microambiente tumorale, la valutazione di biomarcatori con finalità prognostiche e predittive e di strategie per il follow-up e la scoperta di farmaci [3,4,5,6].

Per comprendere meglio il ruolo attuale e le prospettive future dell'IA, è necessario chiarire due termini/definizioni importanti, strettamente associati all'IA: machine learning e deep learning. L'apprendimento automatico è un concetto generale che indica la capacità di una macchina di apprendere e quindi migliorare schemi e modelli di analisi, mentre il deep learning indica un metodo di apprendimento automatico che utilizza reti complesse e profonde per ottenere effetti altamente predittivi [3, 4]. Questi due concetti sono centrali anche nella rivoluzione dell'IA nella gestione dei pazienti oncologici.

Attraverso un approccio basato su una revisione sistematica, ci proponiamo di chiarire quali sono le attuali applicazioni dell'IA in campo oncologico, con un focus specifico sui dispositivi già approvati. Questo approccio consentirà di comprendere meglio i ruoli e le potenzialità dell'IA nella gestione dei pazienti oncologici, rappresentando anche un punto di partenza affidabile per discutere le più importanti prospettive future dell'IA in questo campo.

2. METODI

L'approccio basato sulla revisione sistematica ha aderito al protocollo preimpostato dalla dichiarazione PRISMA [7]. Per fornire un quadro completo della situazione attuale del ruolo svolto dall'IA nella gestione dei pazienti oncologici, è stata eseguita una revisione sistematica, esaminando i dispositivi basati sull'IA che hanno già ottenuto un'approvazione ufficiale per entrare nella pratica clinica oncologica e nei settori correlati. A questo scopo, due autori (C.L. e A.P.) hanno recuperato tutti i dispositivi basati sull'IA che hanno ottenuto l'approvazione della Federal Drug Administration (FDA) in ambito oncologico, estraendo tutti i dati potenziali attraverso la ricerca nei database ufficiali della FDA (https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm514737.pdf; https://www.fda.gov/media/145022/download; https://www.accessda-

ta.fda.gov/scripts/cdrh/cfdocs/cfPMN/denovo.cfm; https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMA/pma.cfm; https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm). Ultimo accesso per tutti i documenti: 05/31/2021. Tali dati sono stati inoltre integrati con tutte le precedenti recensioni o commenti correlati. Tutti i dati sono stati organizzati per essere presentati separatamente per aree oncologiche specifiche nel testo e in una figura riassuntiva (Fig. 1).

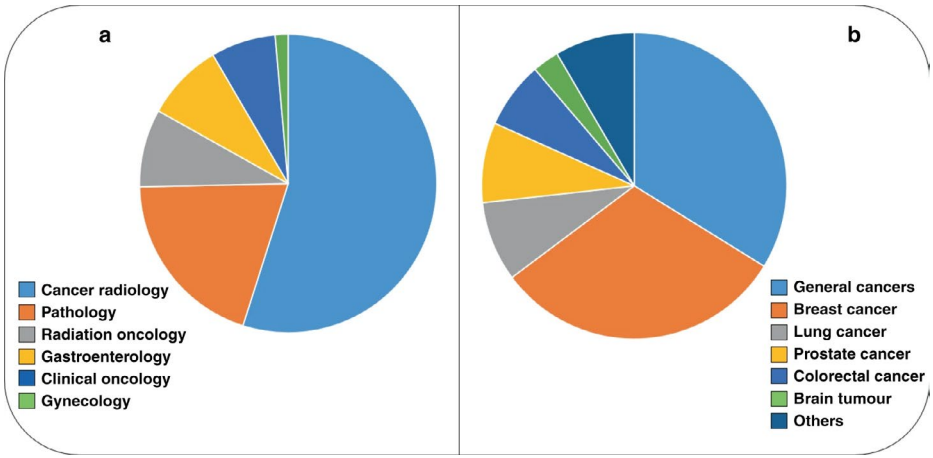


Fig. 1. Stato attuale dell'intelligenza artificiale in oncologia e nei campi correlati. Rappresentazioni sintetiche dei dispositivi basati sull'intelligenza artificiale, approvati dalla FDA, espresse per specialità oncologiche (a: radiologia oncologica 54,9%, patologia 19,7%, radioterapia 8,5%, gastroenterologia 8,5%, oncologia clinica 7,0% e ginecologia 1,4%) e per tipi di tumore (b: tumori generali 33,8%, tumore al seno 31,0%, tumore al polmone 8,5%, tumore alla prostata 8,5%, tumore al colon-retto 7,0% e tumori cerebrali 2,8%, altri: 6 tipi di tumore, 1,4% ciascuno).

3. RISULTATI

Complessivamente, la ricerca ha documentato la presenza di 71 dispositivi associati o associabili all'IA che hanno già ricevuto un'approvazione ufficiale da parte dell'FDA (Tabella 1), corri-

spondendo ai dati di precedenti revisioni correlate [2, 8,9,10]. Il settore oncologico che conta il maggior numero di dispositivi AI è la radiologia oncologica, dove vi sono la maggior parte dei dispositivi approvati (54,9%). Seguono la patologia (19,7%), la radioterapia (8,5%), la gastroenterologia (8,5%), l'oncologia clinica (7,0%) e la ginecologia 1 (1,4%) (Tabella 1, Fig. 1a). La stragrande maggioranza dei dispositivi approvati (>80%) riguardava il complesso settore della diagnostica oncologica.

N°	Data di approvazione	Nome del dispositivo	Descrizione del dispositivo e del suo ruolo	Area di interesse specifica
1	Febbraio 2015	ER APP, Cancro al seno (Visiopharm A/S)	Determinazione della positività e della negatività dei recettori estrogenici nel cancro al seno	Patologia
2	Febbraio 2015	PR APP, Cancro al seno (Visiopharm A/S)	Determinazione della positività e della negatività del recettore del progesterone nel cancro al seno	Patologia
3	Agosto 2015	Kaiku Salute (Kaiku Oy)	Monitoraggio dei risultati e monitoraggio dei sintomi per i pazienti oncologici	Oncologia clinica
4	Novembre 2015	ClearRead CT (Riverain Technologies LLC.)	Assistenza nella revisione degli esami di tomografia computerizzata multi-slice del torace e individuazione di potenziali noduli che il radiologo deve esaminare	Radiologia del cancro
5	Dicembre 2015	Transpara (ScreenPoint Medical BV)	Un ausilio alla lettura per i medici che interpretano le mammografie di screening per identificare le regioni sospette per il cancro al seno	Radiologia del cancro
6	Giugno 2016	SmartTarget (SmartTarget Ltd.)	Procedure interventistiche e diagnostiche guidate da immagini che coinvolgono la ghiandola prostatica	Radiologia del cancro
7	Settembre 2016	Sistema di pianificazione del trattamento Eclipse V15.6 (arian Medical Systems Inc.)	Pianificazione del trattamento radioterapico per pazienti con patologie maligne o benigne	Oncologia radiologica

8	Agosto 2016	LungQ (Thirona Corp.)	Supporto nella diagnosi e nella documentazione di immagini di tessuti polmonari (ad esempio, anomalie) da set di dati CT toracici	Radiologia del cancro
9	Marzo 2017	ColonFlag (Medial EarlySign Inc.)	Supporto per la diagnosi del cancro colorettale ad alto rischio per i pazienti pre-sintomatici	Gastroenterologia
10	Maggio 2017	AmCAD-US (AmCad BioMed Corporation)	Un software per visualizzare e quantificare i dati delle immagini ecografiche con segnali retrodiffusi.	Radiologia del cancro
11	Giugno 2017	C the Signs (C the Signs Ltd.)	Valutazione dei sintomi a supporto della diagnosi del cancro	Oncologia clinica
12	Luglio 2017	QuantX (Approfondimenti quantitativi)	Un sistema di diagnosi dotato di intelligenza artificiale per favorire una diagnosi accurata del cancro al seno.	Radiologia del cancro
13	Dicembre 2017	Veye Chest (Aidence BV)	Supporto per il rilevamento dei noduli polmonari dalla TAC	Radiologia del cancro
14	Gennaio 2018	Arterys Oncology DL (Arterys)	Un software di imaging medicale basato sull'intelligenza artificiale e sul cloud che misura e traccia automaticamente lesioni e noduli nella risonanza magnetica e nella tomografia computerizzata.	Radiologia del cancro
15	Gennaio 2018	GI Genius (Medtronic Inc. (società madre: Medtronic plc.))	Supporto per la diagnosi del cancro colorettale	Gastroenterologia
16	Gennaio 2018	QVCAD (QView Medical Inc.)	Aiuta a rilevare lesioni occulte alla mammografia in regioni non note per i risultati sospetti.	Radiologia del cancro
17	Febbraio 2018	DLCExpert (Mirada Medical Ltd.)	Assistenza al contornamento per la radioterapia da scansioni CT	Oncologia radiologica

18	Maggio 2018	HealthMammo (Zebra Medical Vision Inc.)	Mammografie elaborate e analizzate per lesioni sospette di cancro al seno	Radiologia del cancro
19	Luglio 2018	Arterys Oncology DL (Arterys Inc.)	Supporta il flusso di lavoro oncologico aiutando l'utente a confermare l'assenza o la presenza di lesioni; l'applicazione supporta set di dati anatomici, come TC o MR	Radiologia del cancro
20	Ottobre 2018	Hot Spot APP (Visiopharm A/S)	Metodo di punteggio hotspot per varie applicazioni oncologiche	Patologia
21	Ottobre 2018	APP per il rilevamento dei tumori invasivi (Visiopharm A/S)	Valutazione dei marcatori citocheratina e p63 per la distinzione dei tumori invasivi e non invasivi	Patologia
22	Ottobre 2018	AmCAD-UT (AmCad BioMed Corporation)	Assistenza nell'analisi delle immagini ecografiche della tiroide	Radiologia del cancro
23	Ottobre 2018	Mia - Valutazione intelligente della mammografia (KheironMedical Technologies Ltd.)	Supporto per la diagnosi del cancro al seno a partire dalla mammografia	Radiologia del cancro
24	Ottobre 2018	Arterys MICA (Arterys)	Una piattaforma basata sull'intelligenza artificiale per l'analisi di immagini mediche come la risonanza magnetica e la TAC.	Radiologia del cancro
25	Novembre 2018	SubtlePET (Subtle Medical)	Una tecnologia basata sull'intelligenza artificiale che consente ai centri di offrire una scansione dei pazienti più rapida e sicura, migliorando al contempo la produttività degli esami e la redditività dei fornitori.	Radiologia del cancro
26	Febbraio 2019	DERM (Skin Analytics Ltd.)	Supporto per la diagnosi del cancro della pelle	Oncologia clinica
27	Febbraio 2019	ART-Plan.annotate (heraPanacea SAS)	Contornare il tumore e gli organi circostanti per la radioterapia	Oncologia radiologica

28	Marzo 2019	cmTriage (CureMetrix)	Un software di triage basato sull'intelligenza artificiale per la mammografia.	Radiologia del cancro
29	Aprile 2019	Ricostruzione di immagini con apprendimento profondo (GE Medical Systems)	Una tecnologia di ricostruzione di immagini TC basata sull'apprendimento profondo.	Radiologia del cancro
30	Aprile 2019	Rilevamento automatico dei noduli polmonari (Samsung Electronics Co. Ltd. (società produttrice: Gruppo Samsung)	Rilevamento dei noduli polmonari come supporto diagnostico dalle immagini radiografiche	Radiologia del cancro
31	Maggio 2019	JPC-01K (JLK Inspection Inc.)	Rilevamento del cancro alla prostata come supporto diagnostico dalle immagini di risonanza magnetica	Radiologia del cancro
32	Maggio 2019	syngo.Breast Care (Siemens Healthcare GmbH (società madre: Siemens AG))	Lettura e refertazione per il supporto diagnostico delle mammografie	Radiologia del cancro
33	Giugno 2019	Aquilion ONE (TSX-305A/6) V8.9 con AiCE (Canon MedicalSystems Corporation)	Dispositivo per l'acquisizione e la visualizzazione di volumi in sezione trasversale di tutto il corpo, compresa la testa, con la possibilità di acquisire immagini di interi organi in un'unica rotazione.	Radiologia del cancro
34	Luglio 2019	ProFound AI per la mammografia 2D (iCAD Inc.)	Assistenza per il rilevamento del cancro al seno e soluzione per il flusso di lavoro a partire da mammografie 2D	Radiologia del cancro
35	Luglio 2019	ProFound AI per la tomosintesi mammaria digitale (iCAD Inc.)	Dispositivo software di rilevazione e diagnosi assistita da computer (CAD) da utilizzare durante la lettura degli esami di tomosintesi mammaria digitale (DBT)	Radiologia del cancro
36	Luglio 2019	RayCare 2.3 (RaySearch Laboratories)	Un sistema informativo oncologico utilizzato per supportare i flussi di lavoro, la programmazione e la gestione delle informazioni cliniche per le cure oncologiche e il follow-up.	Radiologia del cancro

37	Agosto 2019	Trattamento radioterapico Ethos (Varian Medical Systems Inc.)	Gestione e monitoraggio dei piani e delle sessioni di trattamento radioterapico	Oncologia radiologica
38	Settembre 2019	AVEC (Valutazione visiva automatizzata della cervice) (MobileODT Ltd.)	Supporto per lo screening del cancro al collo dell'utero per il supporto diagnostico	Ginecologia
39	Settembre 2019	Breast-SlimView (Hera-MI SAS)	Individuazione del cancro al seno per il supporto diagnostico della mammografia	Radiologia del cancro
40	Settembre 2019	Vara (Merantix Healthcare GmbH)	Supporto per lo screening del cancro al seno e triaging dalle mammografie	Radiologia del cancro
41	Ottobre 2019	Software ProFound AI V2.1 (iCAD)	Un dispositivo software CAD destinato ad essere utilizzato dai medici che lo usano durante la lettura della DBT.	Radiologia del cancro
42	Ottobre 2019	DeepDx-Prostate Connect (Deep Bio Inc.)	Riconoscimento dell'adenocarcinoma acinare della prostata	Patologia
43	Novembre 2019	Paige Prostate (Paige Inc.)	Rilevamento del cancro nelle agobiopsie prostatiche	Patologia
44	Novembre 2019	Paige Insight (Paige Inc.)	Visualizzatore di patologia digitale per il supporto diagnostico	Patologia
45	Dicembre 2019	Transpara (ScreenPoint Medical)	Un dispositivo da utilizzare come ausilio alla lettura simultanea per i medici che interpretano mammografie di screening da sistemi FFDM compatibili per identificare le regioni sospette per il cancro al seno e valutarne la probabilità di malignità.	Radiologia del cancro
46	Dicembre 2019	Software QyScore (Qynapse SAS)	Etichettatura automatica, visualizzazione e quantificazione volumetrica di strutture e lesioni cerebrali segmentabili da immagini MR	Radiologia del cancro

47	Dicembre 2019	Discovery AI (Pentax Medical GmbH (società produttrice: Pentax Corporation))	Supporto per il rilevamento dei polipi durante l'esame colorettale	Gastroenterologia
48	Dicembre 2019	RayStation (RaySearch Laboratories AB)	Pianificazione del trattamento e analisi della radioterapia	Oncologia radiologica
49	Dicembre 2019	RayCare 2.3 (RaySearch Laboratories AB)	Supporto ai flussi di lavoro, alla programmazione e alla gestione delle informazioni cliniche per l'assistenza e il follow-up in oncologia.	Oncologia clinica
50	Gennaio 2020	JBD-01K (JLK Inspection Inc.)	Individuazione del cancro al seno per il supporto diagnostico della mammografia	Radiologia del cancro
51	Gennaio 2020	AI-Pathway Companion Prostate Cancer (Siemens Healthcare GmbH (società produttrice: Siemens AG)	Rilevamento del cancro alla prostata per il supporto diagnostico	Oncologia clinica
52	Gennaio 2020	MRCAT Brain (Philips Medical Systems MR Finland (società produttrice: Philips NV))	Pianificazione della radioterapia attraverso la segmentazione automatica delle immagini per i pazienti affetti da tumore cerebrale	Oncologia radiologica
53	Febbraio 2020	InferRead CT Lung (Beijing Infervision Technology Co. Ltd.)	Strumento di screening e gestione del cancro al polmone a partire dalla TAC	Radiologia del cancro
54	Febbraio 2020	b-box (b-rayZ GmbH)	Valutazione della qualità dell'immagine mammografica e della densità del seno dalle mammografie	Radiologia del cancro
55	Febbraio 2020	App per il rilevamento delle metastasi (Visiopharm A/S)	Rilevamento delle metastasi nei linfonodi per l'adenocarcinoma del colon-retto e della mammella	Patologia
56	Febbraio 2020	Galen Prostate (Ibex Medical Analytics Ltd)	Identificazione di un sospetto tumore nelle agobiopsie prostatiche	Patologia
57	Febbraio 2020	densitasAI (Densitas Inc.)	Supporto per la valutazione della densità mammaria dalle mammografie	Radiologia del cancro

58	Marzo 2020	Broncholab (Fluidda Inc)	Supporto nella diagnosi e nella documentazione di immagini del tessuto polmonare (ad esempio, anomalie) da set di dati CT toracici	Radiologia del cancro
59	Marzo 2020	Syngo.CT Lung CAD (Siemens Medical Solutions Inc. (società produttrice: Siemens AG))	Assistenza nell'individuazione di noduli polmonari solidi durante la revisione degli esami di tomografia computerizzata multidetettore del torace	Radiologia del cancro
60	Marzo 2020	MammoScreen (Therapixel SA)	Aiuta a identificare i reperti su FFDM di screening acquisiti con sistemi mammografici compatibili e valuta il livello di sospetto	Radiologia del cancro
61	Marzo 2020	CAD EYE (FUJIFILM Europe GmbH)	Supporto per il rilevamento e la caratterizzazione dei polipi del colon durante la colonscopia	Gastroenterologia
62	Maggio 2020	Sistema endoscopico a capsula NaviCam con capsula per lo stomaco NaviCam (AnX Robotica, Inc.).	È un sistema di endoscopia a capsula manovrata magneticamente; è costituito da una capsula ingeribile e da un controller magnetico e viene utilizzato per la visualizzazione dello stomaco e del duodeno. Il controller magnetico viene utilizzato all'esterno del paziente ed è accoppiato magneticamente alla capsula per controllarne la posizione e la direzione di visualizzazione.	Gastroenterologia
63	Giugno 2020	Test di mutazione Cobas® EZH2 (Roche Molecular System, Inc.)	Il test è destinato all'identificazione dei pazienti affetti da linfoma follicolare con una mutazione EZH2 per il trattamento con TAZVERIK (tazemetostat); in abbinamento all'analizzatore cobas z 480.	Patologia

64	Luglio 2020	Cocktail di sonde dna Her2 dual ish	È finalizzato a determinare lo stato di amplificazione del gene HER2 mediante il calcolo del rapporto tra il gene HER2 e il cromosoma 17 al microscopio ottico.	Patologia
65	Ottobre 2020	Citologia Cintec plus (Ventana Medical Systems, Inc.)	Test immunocitochimico qualitativo per la rilevazione simultanea delle proteine p16INK4a e Ki-67 in campioni cervicali, destinato alla diagnosi del cancro cervicale.	Patologia
66	Novembre 2020	Genius AI Detection (Hologic, Inc.)	Dispositivo software per l'identificazione di potenziali anomalie nelle immagini di tomosintesi mammaria	Radiologia del cancro
67	Novembre 2020	FoundationOne Liquid CDx (Foundation Medicine, Inc.)	Si tratta di un test qualitativo basato sulla NGS che interroga 311 geni. Utilizza il DNA libero da cellule circolanti (cfDNA) isolato dal plasma di pazienti affetti da cancro ed è destinato a essere utilizzato come diagnostico di accompagnamento per identificare i pazienti che possono beneficiare del trattamento con terapie mirate (bersagli identificati con NGS).	Patologia
68	Gennaio 2021	Densità mammaria Visage (Visage Imaging)	L'applicazione software è destinata all'uso con la mammografia digitale full-field compatibile, per aiutare i radiologi nella valutazione della composizione del tessuto mammario.	Radiologia del cancro
69	Gennaio 2021	Sistema di imaging mammario Imagio (Seno Medical Instruments, Inc.)	Consente una migliore classificazione delle masse mammarie rispetto alla sola ecografia; include un software basato sull'intelligenza artificiale.	Radiologia del cancro

70	Aprile 2021	Pannello MMR RxDx VENTANA (Ventana Medical Systems, Inc.)	CDx per identificare le pazienti con carcinoma endometriale con stato dMMR che possono beneficiare del trattamento con Jemperli (dostarlimab-gxly).	Patologia
71	Aprile 2021	GI Genius (Cosmo Artificial Intelligence-AI, LTD)	Si tratta di uno strumento di lettura assistita dal computer progettato per aiutare gli endoscopisti a rilevare le lesioni della mucosa del colon (come polipi e adenomi) in tempo reale durante l'endoscopia standard a luce bianca.	Gastroenterologia

Tabella 1 Elenco dei dispositivi medici dotati di AI/associati approvati dalla FDA statunitense specificamente per i settori oncologici.

- *Sintesi delle diverse aree mediche correlate all'oncologia di tutti i dispositivi associati all'IA approvati dalla FDA: 39 radiologia oncologica (54,9%); 14 patologia (19,7%); 6 radioterapia oncologica (8,5%); 6 gastroenterologia (8,5%); 5 oncologia clinica (7,0%), ginecologia 1 (1,4%).*
- *Sintesi dei diversi tipi di tumore studiati dai dispositivi presentati: 24 tumori generali (33,8%); 22 tumori al seno (31,0%); 6 tumori al polmone (8,5%); 6 tumori alla prostata (8,5%); 5 tumori al colon-retto (7,0%); 2 tumori al cervello (2,8%); 6 altri (6 tipi): 1,4% ciascuno.*
- *AI intelligenza artificiale, US FDA United States Food and Drug Administration, CT tomografia computerizzata, MRI risonanza magnetica, ECG elettrocardiogramma, CAD rilevamento/diagnosi assistita da computer, DBT tomosintesi mammaria digitale, FFDM mammografia digitale a tutto campo.*

Per quanto riguarda i diversi tipi di tumore che possono essere studiati adottando tali dispositivi, la maggior parte di essi è stata concepita per essere applicata a un ampio spettro di tumori solidi maligni (cancro in generale, 33,8%). Il tumore specifico che conta il maggior numero di dispositivi AI è il tumore al seno (31,0%), seguito dal tumore al polmone e alla prostata (8,5% ciascuno), dal tumore del colon-retto (7,0%), dai tumori cerebrali (2,8%) e da altri (6 tipi, 1,4% ciascuno) (Tabella 1, Fig. 1b).

4. DISCUSSIONE E PROSPETTIVE FUTURE

In questo articolo viene fornita una panoramica completa delle attuali applicazioni dell'IA in ambito oncologico, descrivendo in particolare i dispositivi basati sull'IA che hanno già ottenuto l'approvazione ufficiale per la pratica clinica. Fin dalla sua nascita, l'IA ha dimostrato la sua importanza trasversale in tutte le branche scientifiche, mostrando un impressionante potenziale di crescita per il futuro. Come evidenziato in questo studio, questa crescita ha interessato anche l'oncologia e le specialità correlate.

In generale, l'applicazione dei dispositivi approvati dalla FDA non è stata concepita come un sostituto del classico flusso di lavoro analitico/diagnostico, ma è intesa come uno strumento integrativo, da utilizzare in casi selezionati, che può rappresentare il passo decisivo per migliorare la gestione dei pazienti oncologici. Attualmente, in questo campo, i settori in cui l'IA sta avendo un impatto maggiore sono rappresentati dalle aree diagnostiche, che contano per la stragrande maggioranza dei dispositivi approvati (>80%), e in particolare la radiologia e la patologia.

La diagnostica del cancro rappresenta classicamente il punto di partenza necessario per la progettazione di approcci terapeutici e gestione clinica appropriati, e il suo perfezionamento basato sull'IA è un risultato molto importante. Inoltre, ciò indica che gli sviluppi futuri dell'IA dovrebbero considerare anche orizzonti inesplorati ma cruciali in questo panorama, tra cui la scoperta di farmaci, la somministrazione della terapia e le strategie di follow-up. A nostro avviso, per determinare un miglioramento decisivo nella gestione dei pazienti oncologici, la crescita dell'IA dovrebbe seguire modelli completi e multidisciplinari. Questa rappresenta una delle più importanti opportunità offerte dall'IA, che consentirà la corretta interazione e integrazione delle aree oncologiche su uno specifico paziente, rendendo possibili gli impegnativi obiettivi della medicina personalizzata.

I tipi di tumore specifici che ora stanno sperimentando maggiori vantaggi dai dispositivi basati sull'intelligenza artificiale

nella pratica clinica sono innanzitutto il tumore al seno, il tumore al polmone e il tumore alla prostata. Questo dovrebbe essere visto come il riflesso diretto della loro maggiore incidenza rispetto ad altri tipi di tumore, ma in futuro dovrebbero essere presi in considerazione altri tipi di tumore, compresi quelli rari che ancora soffrono della mancanza di approcci standardizzati. Poiché l'IA si basa sulla raccolta e sull'analisi di ampie serie di casi, tuttavia, il miglioramento del trattamento delle neoplasie rare rappresenterà probabilmente un risultato tardivo. In particolare, se considerati insieme, i tumori rari sono una delle categorie più importanti dell'oncologia di precisione [11]. Pertanto, a nostro avviso, le strategie di sviluppo dell'IA in corso non possono ignorare questo gruppo di tumori; anche se i potenziali benefici sembrano lontani, è già tempo di iniziare a raccogliere dati sulle neoplasie rare.

Una delle aspettative più promettenti per l'IA è la possibilità di integrare dati diversi e compositi derivati da approcci multiomici ai pazienti oncologici. I promettenti strumenti dell'IA potrebbero essere gli unici in grado di gestire la grande quantità di dati provenienti da diversi tipi di analisi, comprese le informazioni derivate dal sequenziamento di DNA e RNA. Su questa linea, la recente pubblicazione degli standard e delle linee guida dell'American College of Medical Genetics per l'interpretazione delle varianti di sequenza [12] ha favorito una nuova ondata di sviluppo dell'IA, con opportunità innovative nell'oncologia di precisione (https://www.businesswire.com/news/home/20190401005976/en/Fabric-Genomics-Announces-AI-based-ACMG-Classification-Solution-for-Genetic-Testing-with-Hereditary-Panels; ultimo accesso 21/09/2021). A nostro avviso, tuttavia, la mancanza di informazioni veritiere derivate da archivi di dati sanitari protetti rappresenta ancora un ostacolo nella valutazione dell'accuratezza delle applicazioni di IA per il processo decisionale clinico.

Nel complesso, l'IA sta avendo un impatto crescente su tutte le branche scientifiche, compresa l'oncologia e i campi ad essa correlati, come evidenziato in questo studio. Per progettare nuo-

ve strategie di sviluppo con impatti concreti, i primi passi sono rappresentati dalla conoscenza del suo background storico e dalla comprensione dei suoi attuali risultati. Come qui evidenziato, l'IA è già entrata nella pratica clinica oncologica, ma è necessario un impegno continuo e crescente per consentire all'IA di esprimere tutto il suo potenziale. A nostro avviso, in questo momento le sfide più importanti per portare a termine la "rivoluzione dell'IA" in oncologia sono la creazione di una visione multidisciplinare/integrativa dello sviluppo, l'immediata comprensione dell'importanza di tutte le neoplasie, compresi i tumori rari, e il continuo sostegno per garantirne la crescita.

9 - INTELLIGENZA ARTIFICIALE E VISIONE COMPUTERIZZATA NELLA LOMBALGIA: UNA REVISIONE SISTEMATICA

Tratto e tradotto da
D'Antoni F, Russo F, Ambrosio L, Vollero L, Vadalà G, Merone M, Papalia R, Denaro V. *Artificial Intelligence and Computer Vision in Low Back Pain: A Systematic Review.* Int J Environ Res Public Health. 2021 Oct 17;18(20):10909.

doi: 10.3390/ijerph182010909. PMID: 34682647; PMCID: PMC8535895.

Le parti omesse dal curatore rispetto all'originale sono indicate dal segno **[...]**

ABSTRACT

Il dolore lombare cronico (LBP [*Low Back Pain*]) è un sintomo che può essere causato da diverse patologie ed è attualmente la principale causa di disabilità a livello mondiale. La crescente quantità di immagini digitali in ortopedia ha portato allo sviluppo di metodi legati all'intelligenza artificiale, e alla computer vision in particolare, che mirano a migliorare la diagnosi e il trattamento del LBP. In questa revisione abbiamo esaminato sistematicamente la letteratura disponibile sull'uso della computer vision nella diagnosi e nel trattamento del LBP. È stata effettuata una ricerca sistematica nel database elettronico PubMed. La strategia di ricerca è stata impostata sulla combinazione delle seguenti parole chiave: "Intelligenza artificiale", "Estrazione di caratteristiche", "Segmentazione", "Computer Vision", "Machine Learning", "Deep Learning", "Neural Network", "Low Back

Pain", "Lumbar". Risultati: la ricerca ha prodotto un totale di 558 articoli. Dopo un'attenta valutazione degli abstract, ne sono stati esclusi 358, mentre 124 articoli sono stati esclusi dopo l'esame dell'intero testo, portando il numero di articoli selezionabili a 76. Le principali applicazioni della computer vision in LBP includono l'estrazione e la segmentazione delle caratteristiche, che di solito sono seguite da ulteriori compiti. La maggior parte dei metodi recenti utilizza modelli di deep learning piuttosto che tecniche di elaborazione digitale delle immagini. I metodi più performanti per la segmentazione di vertebre, dischi intervertebrali, canale spinale e muscoli lombari raggiungono punteggi Sørensen-Dice superiori al 90%, mentre gli studi incentrati sulla localizzazione e l'identificazione delle strutture hanno mostrato complessivamente un'accuratezza superiore all'80%. Si prevede che i futuri progressi dell'intelligenza artificiale aumenteranno l'autonomia e l'affidabilità dei sistemi, fornendo così strumenti ancora più efficaci per la diagnosi e il trattamento del LBP.

Parole chiave: lombalgia, ortopedia, intelligenza artificiale, computer vision, elaborazione digitale delle immagini, deep learning, sistemi di supporto alle decisioni, diagnosi assistita da computer

1. INTRODUZIONE

Nell'ultimo decennio si è assistito a un aumento significativo dell'uso dell'intelligenza artificiale (IA) nei campi più disparati, dagli assistenti vocali comunemente impiegati nella vita quotidiana alle automobili a guida autonoma. Grazie alla capacità unica delle macchine intelligenti di essere addestrate e di imparare automamente nuovi compiti sulla base di esperienze precedenti o di dati forniti, l'uso dell'IA è sempre più studiato per applicazioni nella ricerca medica [1]. In effetti, i computer basati sull'IA hanno già dimostrato di poter rivoluzionare la progettazione e la scoperta di farmaci [2,3], la segmentazione automatica e l'estra-

zione di dati rilevanti da insiemi di dati radiologici [4], nonché la formulazione di diagnosi, la previsione di esiti e la pianificazione di trattamenti in diversi campi medici [5,6,7]. L'adozione di questa tecnologia innovativa viene esplorata anche nella chirurgia vertebrale [1]. Infatti, grazie alla sua natura interdisciplinare e all'ampio utilizzo di immagini radiologiche per ispezionare le strutture anatomiche della colonna vertebrale, l'uso dell'IA può essere di particolare utilità per determinare, ad esempio, quali sono i dischi problematici [8], classificare una curva scoliotica [9] e prevederne la progressione [10]. In questo studio abbiamo esaminato sistematicamente la letteratura disponibile sull'uso dell'IA, e più specificamente della computer vision, nella prevenzione, nella diagnosi e nel trattamento della lombalgia cronica (LBP).

Il LBP è causato principalmente dalla degenerazione del disco intervertebrale ed è attualmente la principale causa di disabilità in tutto il mondo, nonché il motivo più comune per le richieste di risarcimento dei lavoratori [11]. L'IA ha migliorato la pratica clinica per quanto riguarda il trattamento, la prevenzione e la previsione degli esiti dei soggetti affetti da LBP. Ciò è dovuto principalmente alla quantità sempre crescente di dati clinici a disposizione degli operatori, che consentono di addestrare e sviluppare metodologie di IA sempre più sofisticate. Per quanto riguarda in particolare il LBP, ogni giorno viene raccolta un'enorme quantità di immagini cliniche digitali per rilevare i segni di malattia nelle strutture spinali. Per questo motivo, negli ultimi anni sono stati sviluppati diversi algoritmi di apprendimento automatico per accelerare il processo diagnostico e ottimizzare il recupero dei pazienti. I più recenti miglioramenti dell'IA sono stati accompagnati dall'esplosione del deep learning e dall'aumento della capacità di calcolo, che consentono di sviluppare modelli sempre più autonomi e accurati. In particolare, le tecniche di computer vision applicate alle immagini cliniche consentono di rilevare alcune caratteristiche dell'immagine invisibili all'occhio umano. La computer vision in relazione alla LBP è importante sotto molti aspetti: permette di eseguire diversi compiti che possono migliorare la pratica clinica, come la localizzazione

e il rilevamento automatico delle strutture lombari con segmentazione. Inoltre, consente di estrarre dall'immagine un insieme di caratteristiche che possono essere utilizzate come input per ulteriori algoritmi di apprendimento automatico al fine di fornire un supporto decisionale al medico o, in altri casi, suggerire direttamente la diagnosi più appropriata. Per questo motivo, abbiamo rivisto sistematicamente la letteratura disponibile sull'applicazione della computer vision alla diagnosi e al trattamento del LBP, al fine di descrivere lo stato dell'arte di tale tecnologia e le sue potenziali applicazioni.

2. MATERIALI E METODI

Per effettuare una ricerca esaustiva degli articoli di IA relativi al LBP, abbiamo effettuato una ricerca su PubMed (query di ricerca utilizzata: (((Intelligenza artificiale [Titolo/Abstract]) OPPURE ((estrazione di caratteristiche[Titolo/Abstract]) OPPURE ((segmentazione[Titolo/Abstract]) OPPURE (Computer Vision[Titolo/Abstract]) OPPURE (Machine learning[Titolo/Abstract])) OR (deep learning[Titolo/Abstract]) OR (rete neurale[Titolo/Abstract]))) AND ((Low Back Pain [Titolo/Abstract]) OR (lumbar[Title/Abstract]))). Tutte le parole di ricerca dovevano essere incluse nel titolo o nell'abstract degli articoli: i termini "low back pain" e "lumbar" sono stati considerati per la parte patologica, mentre i termini "artificial intelligence", "feature extraction", "segmentation", "computer vision", "machine learning", "deep learning" e "neural network" sono stati considerati per la parte AI. Abbiamo selezionato tutti gli articoli che includevano almeno un termine della parte patologica e almeno un termine della parte di intelligenza artificiale nel titolo o nell'abstract.

2.1 CRITERI DI INCLUSIONE E DI ESCLUSIONE

Lo scopo di questo lavoro è stato quello di raccogliere tutti i lavori riguardanti l'utilizzo dell'IA, e in particolare della compu-

ter vision, nella diagnosi, nella prevenzione e nel trattamento del LBP cronico e delle patologie correlate. Tutti gli articoli selezionati dovevano soddisfare i seguenti criteri di inclusione:

- Il LBP cronico o le patologie lombari dovevano essere tra gli argomenti principali dell'articolo. Sono stati inclusi lavori sulla prevenzione, la diagnosi o il trattamento del LBP cronico e sul trattamento di almeno una delle strutture coinvolte nel LBP (ad esempio, vertebre, dischi, muscoli);
- L'IA deve essere stata utilizzata con applicazione alle immagini cliniche. Sono stati inclusi articoli che sfruttavano metodi di IA nelle aree della computer vision, dell'apprendimento automatico e delle reti neurali artificiali (NN);
- Soggetti dello studio: tutti gli articoli devono essere basati su studi sulla lombalgia umana e sulla relativa patologia, indipendentemente dall'età o dall'occupazione dei soggetti inclusi nello studio;
- Lingua: tutti gli articoli devono essere scritti in inglese.

Al contrario, gli articoli esclusi non soddisfacevano i criteri di inclusione per uno dei seguenti motivi:

- È stato considerato un problema medico diverso: abbiamo escluso gli articoli che non consideravano il LBP cronico e le strutture fisiche e i dati medici correlati. Ad esempio, sono stati esclusi gli studi che consideravano solo le vertebre cervicali o toraciche, o che si concentravano su osteoporosi, metastasi, LBP traumatico e altre cause di LBP non discogenico;
- L'IA non è stata presa in considerazione: alcuni articoli nei risultati della ricerca hanno proposto definizioni e pratiche per il LBP basate solo sull'osservazione medica senza l'utilizzo dell'IA;
- La computer vision e le immagini cliniche non sono state prese in considerazione nello studio, indipendentemente

dal fatto che l'IA fosse utilizzata per sviluppare diagnosi o sistemi di supporto;
- Studi su animali: abbiamo escluso gli studi basati su strutture vertebrali di animali;
- Studi embrionali: sono stati esclusi gli studi effettuati su embrioni e riguardanti l'embriogenesi delle strutture spinali.

Uno screening preliminare della selezione degli articoli ci ha permesso di definire tre categorie principali in cui potrebbe essere suddiviso l'utilizzo dell'IA nella LBP, ovvero la computer vision, la diagnosi assistita da computer e i sistemi di supporto alle decisioni (DSS) (Figura 1). La computer vision è il campo dell'IA che studia come i computer possano ottenere una comprensione di alto livello da immagini o video digitali. Per quanto riguarda l'LBP, le sue applicazioni principali riguardano l'estrazione di caratteristiche e la segmentazione delle immagini. L'estrazione delle caratteristiche è un processo di riduzione della dimensionalità che viene applicato alle immagini ottenute tramite risonanza magnetica (MRI), ultrasuoni, raggi X e tomografia computerizzata (CT). L'obiettivo principale dell'estrazione delle caratteristiche è quello di recuperare un numero limitato di caratteristiche rilevanti da un'immagine senza perdere informazioni importanti, al fine di facilitare compiti successivi come la classificazione o la regressione.

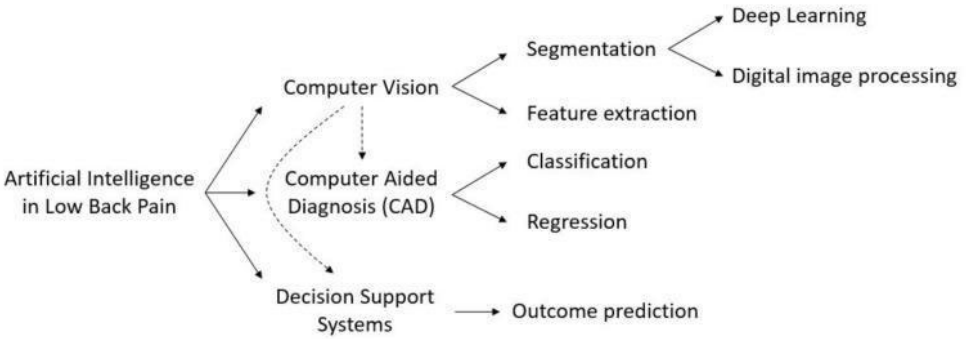


Figura 1. *Suddivisione schematica dei compiti dell'IA applicata a LBP.*

La segmentazione delle immagini consiste nel dividere un'immagine in sottoregioni corrispondenti a diversi elementi dell'immagine. Più in profondità, l'obiettivo della segmentazione delle immagini è l'etichettatura di ogni pixel di un'immagine con una classe corrispondente, ad esempio primo piano o sfondo, al fine di individuare gli elementi rilevanti di un'immagine. Si ricorre principalmente a due tecniche: il deep learning, in cui l'immagine viene data direttamente in input a un NN artificiale che viene addestrato su altre immagini per identificare automaticamente le sottoregioni, e le tecniche di elaborazione digitale delle immagini (DIP), che elaborano le immagini digitali per trovare i bordi delle diverse regioni sulla base di caratteristiche semantiche, sfruttando metodi come il gradient thresholding o i modelli statistici di forma.

La diagnosi assistita dal computer è un gruppo di tecniche che aiutano i medici a identificare una patologia o a quantificare il grado di una malattia. Può essere suddivisa in classificazione e regressione, in cui i modelli di apprendimento automatico o profondo vengono utilizzati rispettivamente per assegnare un'etichetta predefinita o per generare un risultato numerico. In pratica, la classificazione viene utilizzata per identificare o classificare una patologia, mentre la regressione viene utilizzata per produrre una valutazione quantitativa di una certa misura.

I sistemi di supporto alle decisioni (DSS) sono sistemi software che consentono ai medici di migliorare il processo decisionale e l'esito dei pazienti affetti da una specifica malattia. L'obiettivo della stragrande maggioranza dei DSS è la previsione dell'esito, ossia la previsione del miglioramento che un paziente sperimenterà dopo l'esposizione a una determinata terapia. Prevedendo la misura in cui un paziente beneficerebbe di un trattamento specifico, i DSS forniscono al medico strumenti pratici per valutare se l'intervento chirurgico sia preferibile o meno al trattamento conservativo. Infine, i DSS possono essere utilizzati per la prevenzione, ad esempio fornendo all'utente raccomandazioni o pratiche corrette per prevenire l'insorgere di una malattia. Vale la pena notare che le tecniche di computer vision possono essere

utilizzate come pre-elaborazione per lo sviluppo di un DSS, così come per la diagnosi assistita.

2.2 METRICHE DI VALUTAZIONE

Compiti diversi usano metriche diverse per valutare le prestazioni dei sistemi di intelligenza artificiale. Tuttavia, considerando la grande quantità di lavori riportati in questa rassegna, sono state prese in considerazione metriche diverse anche all'interno dello sesso compito. Per quanto riguarda il compito di estrazione delle caratteristiche, non è stata considerata alcuna metrica di valutazione specifica. Ciò è dovuto al fatto che, nella maggior parte dei casi, l'estrazione delle caratteristiche diventa la fase preliminare per ulteriori compiti, come la classificazione e la regressione, e la maggior parte dei lavori riporta solo le prestazioni di quest'ultima.

Per quanto riguarda il compito di classificazione, abbiamo riportato i risultati in termini di accuratezza (Acc), laddove disponibili. Per brevità, consideriamo un compito di classificazione binario, ad esempio positivo o negativo. Dato un insieme di test composto da N campioni, definiti i Veri Positivi (TP) come il numero di campioni positivi classificati correttamente, e i Veri Negativi (TN) come il numero di campioni negativi classificati correttamente, l'accuratezza è definita da:

$$Acc\%=TP+TNN\times100$$

Pertanto, valori maggiori corrispondono a prestazioni migliori. Per ogni classe è possibile calcolare anche il richiamo e la precisione. Definiti i falsi positivi (FP) e falsi negativi (FN) come numero di campioni positivi/negativi erroneamente classificati, il richiamo e la precisione vengono calcolati come segue:

$$Recall=TPTP+FNPrecision=TPTP+FP$$

Nei problemi binari, il richiamo è chiamato anche "tasso di veri positivi" e corrisponde alla sensibilità, mentre il tasso di veri

negativi è chiamato specificità. Nel caso di problemi multiclasse, l'accuratezza viene calcolata considerando la TP per ogni classe e si possono calcolare il richiamo e la precisione per classe. Per i set di dati sbilanciati, il punteggio F1 può essere calcolato per ogni classe. Il punteggio F1 per la classe *c* è così definito:

$$F1\text{-}Score_c = 2\text{-}Recall_c\text{-}Precision_c Recall_c + Precision_c$$

Questo tiene conto sia del richiamo che della precisione della classe. Un'altra metrica di valutazione molto utilizzata è l'area sotto la curva (AUC [*Area Under the Curve*]), che corrisponde all'area sotto la curva ROC (*Receiver Operating Characteristic*) che mostra le prestazioni di un modello di classificazione a tutte le soglie di classificazione, tracciate considerando il tasso di veri positivi contro il tasso di falsi positivi. I suoi valori vanno da 0 a 1 (più si avvicina a 1, migliore è la prestazione).

Per quanto riguarda il compito di regressione, consideriamo una sequenza di valori originali $x(t)$ e una sequenza di valori previsti $x\tilde{}(t)$. L'errore assoluto medio (MAE) per una sequenza di *N* timestamp è definito come:

$$MAE = \sum t = 1N |x(t) - x\tilde{}(t)| N$$

Quindi, più il valore è vicino allo 0, migliore è la prestazione. In alcuni casi, per valutare le prestazioni si utilizzano valori di errore percentuale, il cui significato varia a seconda del compito preso in considerazione.

Per quanto riguarda il compito di segmentazione, vengono utilizzati due principali indici di prestazione percentuale che valutano in che misura il risultato della segmentazione si avvicina alla segmentazione desiderata. Come detto, la segmentazione consiste nell'etichettare ogni pixel di un'immagine. Dati due insiemi di dati *A* e *B*, corrispondenti all'area desiderata e a quella effettivamente segmentata, il coefficiente di Sørensen-Dice (DICE) è definito come:

$$DICE(A,B)=2-|A\cap B||A|+|B|$$

dove $|A|$ e $|B|$ sono le cardinalità dei due insiemi. Divide il numero di elementi comuni dei due insiemi per il numero totale di elementi dei due insiemi. Se applicato a dati binari, è equivalente al *punteggio F1*. Diversamente, l'indice di Jaccard è definito come:

$$Jaccard(A,B)=|A\cap B||A\cup B|$$

ed è noto anche come Intersezione su Unione. Per entrambi gli indici, più ci si avvicina al 100%, migliori sono le prestazioni. Vale la pena notare che

$$DICE(A,B)\geq Jaccard(A,B)$$

per qualsiasi coppia di insiemi

$$(A,B)$$

e la relazione

$$Jaccard=DICE/(2-DICE)$$

servono per calcolare un valore a partire dall'altro.

3. QUALITÀ DELLE PROVE

La qualità metodologica degli studi inclusi è stata valutata in modo indipendente da due revisori (L.A. e F.R.) e ogni disaccordo è stato risolto con l'intervento di un terzo revisore (G.V.). I rischi di distorsione e l'applicabilità degli studi inclusi sono stati valutati utilizzando criteri di valutazione personalizzati basati sul Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) [12]. Questo strumento si basa su 4 domini: selezione del

paziente, test indice, standard di riferimento, flusso e tempistica. Ogni dominio è valutato in termini di rischio di bias e i primi 3 domini sono valutati anche in termini di problemi di applicabilità. Sessantotto studi sono stati valutati su una scala a 3 punti, riflettendo le preoccupazioni relative al rischio di bias e all'applicabilità come basso, non chiaro o alto, come mostrato nella Figura 2 (i dettagli dell'analisi sono presentati nelle Tabelle S1 e S2).

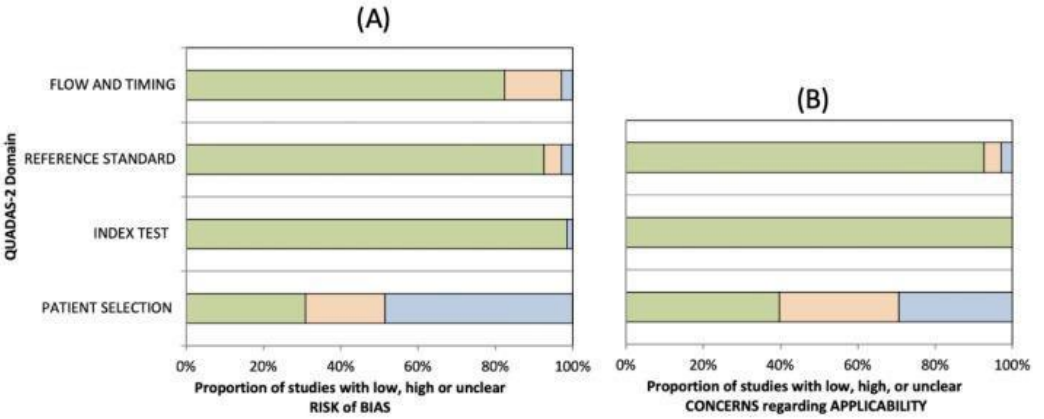


Figura 2. *Riassunto della qualità metodologica degli studi inclusi in relazione ai 4 domini che valutano il rischio di bias (A) e ai 3 domini che valutano i problemi di applicabilità (B) del punteggio QUADAS-2. La parte di studi con un basso rischio di distorsione è evidenziata in verde, gli studi con un rischio di distorsione non chiaro sono rappresentati in blu e gli studi con un alto rischio di distorsione sono rappresentati in arancione.*

4. I RISULTATI

La ricerca è stata effettuata il 18 marzo 2021 e ha prodotto 558 articoli. Tuttavia, molti di questi articoli si concentravano su un argomento diverso da quello di questa revisione, quindi dopo un primo screening basato sui titoli e sugli abstract degli articoli, abbiamo ridotto il numero di articoli ammissibili a 200. Una seconda fase di screening è stata eseguita dopo aver letto il testo completo di ogni articolo, che ha portato il totale degli arti-

coli inclusi a 76. Una seconda fase di screening è stata effettuata dopo aver letto il testo completo di ogni articolo, che ha portato il totale degli articoli inclusi a 76. Abbiamo creato un diagramma di flusso secondo il protocollo PRISMA che mostra il processo di selezione degli studi (Figura 3). Gli articoli sono stati vagliati da due revisori indipendenti che, in caso di discrepanze sull'inclusione o l'esclusione di un articolo, hanno discusso fino a raggiungere un consenso.

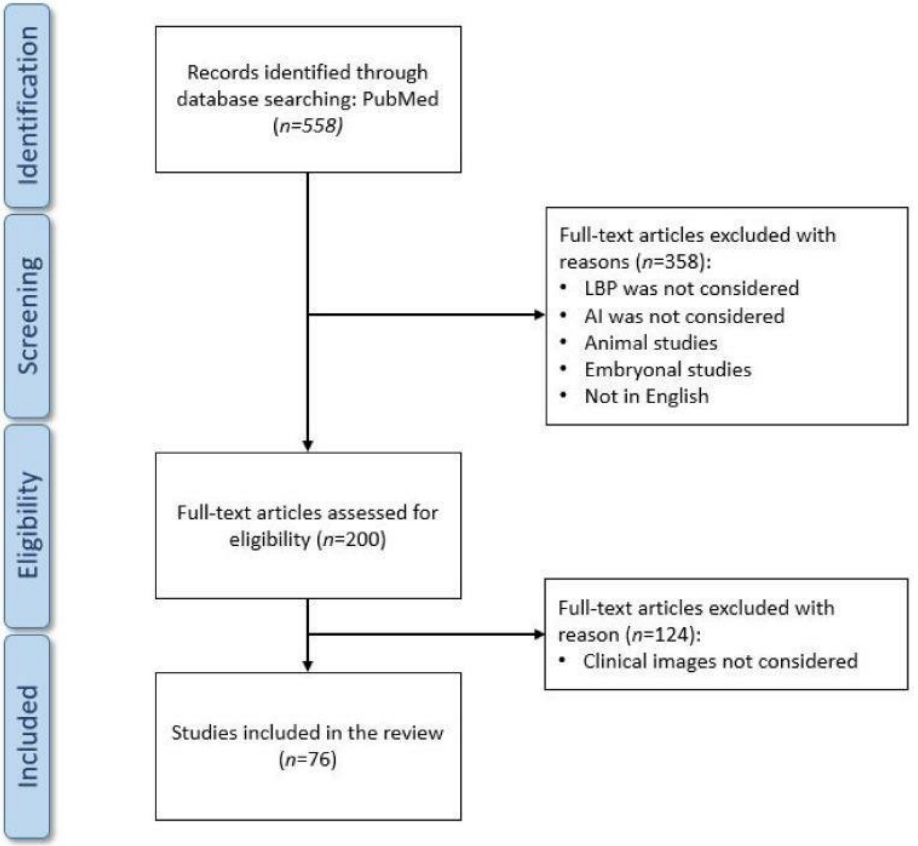


Figura 3. *Diagramma di flusso del Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA).*

Vale la pena notare come la quantità di lavori pubblicati stia aumentando di anno in anno e che il numero di articoli pubblicati nel 2020 è quasi doppio rispetto a quello del 2019. Ciò può essere dovuto a due ragioni principali: in primo luogo, la quantità sempre maggiore di immagini e dati clinici a disposizione dei ricercatori e, in secondo luogo, il miglioramento della capacità di calcolo osservato negli ultimi anni. I risultati finali della ricerca comprendono anche cinque revisioni. Una di queste, pubblicata nel 2020 da Tagliaferri et al. [13], è specificamente incentrata sulla LBP, ma considera solo la capacità di diagnosi e prognosi dell'IA rispetto ai metodi McKenzie e STarT Back, senza prendere in considerazione i lavori che sfruttano le immagini cliniche. Le altre quattro rassegne non si concentrano specificamente sulla LBP. Entrando più nei dettagli: nel 2019 Tack [14] si è concentrato sulla medicina muscoloscheletrica in generale, determinando in quali campi l'IA ha raggiunto i livelli di previsione simili a quelli umani; nel 2020 Azimi et al. [15] si sono concentrati sull'uso delle NN per il trattamento dell'intera colonna vertebrale; nel 2019 Galbusera et al. [1] hanno descritto l'applicazione dell'IA a problemi relativi all'intera colonna vertebrale; infine, nel 2016 Yao et al. [16] hanno eseguito uno studio comparativo multicentrico su milestone per i metodi di segmentazione vertebrale basati su immagini TC. Sono stati trovati anche due articoli che presentano banche dati: LUMINOUS, che è un database di immagini ecografiche di 109 pazienti per la segmentazione del muscolo multifido [17], e MyoSegmentum, che include immagini di risonanza magnetica di 54 pazienti per la segmentazione dei muscoli lombari e dei corpi vertebrali [18].

La parte restante di questa sezione riporta i risultati della ricerca sulla computer vision. In particolare, abbiamo elencato i manoscritti che hanno svolto un compito di estrazione di caratteristiche o di segmentazione semantica e abbiamo descritto i lavori che hanno utilizzato approcci DIP/NN in due diverse sottosezioni.

4.1 ESTRAZIONE DELLE CARATTERISTICHE

L'estrazione delle caratteristiche è un processo di riduzione della dimensionalità volto a identificare un insieme ristretto di caratteristiche rilevanti per migliorare la capacità predittiva di un sistema. In questa rassegna, abbiamo identificato un totale di 8 lavori che mirano a estrarre caratteristiche rilevanti da diversi tipi di immagini LBP; le caratteristiche principali di questi studi sono riportate nella Tabella 1. In dettaglio, abbiamo incluso:

- sei articoli sulla risonanza magnetica (1 dei quali prende in considerazione la risonanza magnetica 3D);
- un articolo sulle immagini 3D della superficie posteriore;
- un articolo sull'imaging a raggi X.

Autore/Anno	Compito principale	Tipo di dati	# Pazienti	Strutture coinvolte	Risultati	Modello
Adankon, 2012 [19]	Estrazione e classificazione delle caratteristiche	Immagine 3D della superficie posteriore	165	Vertebre	Acc = 95%	Descrittori geometrici locali e SVM
Castro-Mateos, 2014 [20]	Estrazione di caratteristiche e segmentazione	Risonanza magnetica 3D	59	Dischi	DICE = 88,4%	Spazio del modello statistico di forma e spazio B-Spline
Raudner, 2020 [21]	Estrazione delle caratteristiche	Risonanza magnetica	58	Dischi	/	GRAPPATINI
Abdollah, 2020 [22]	Estrazione delle caratteristiche	Risonanza magnetica	28	Dischi, Vertebre	/	Random Forest e analisi delle texture
Yang 2020 [8]	Estrazione e classificazione delle caratteristiche	Risonanza magnetica	109	Dischi	Acc = 88,3%	Trasformata wavelet di Gabor e tracciamento di feature KLT
Ruiz-España, 2015 [23]	Estrazione e classificazione delle caratteristiche	Risonanza magnetica	67	Dischi	Acc > 90%	Flusso vettoriale di gradienti, diversi modelli di ML

Ketola, 2020 [24]	Estrazione e classificazione delle caratteristiche	Risonanza magnetica	518	LBP	Acc = 83%	Estrazione delle caratteristiche della texture e regressione logistica
Garcia-Cano, 2018 [10]	Estrazione di caratteristiche e regressione	Raggi X	150	Vertebre	Angolo di Cobb MAE = 4,79°	Analisi delle componenti indipendenti e Random Forest

Tabella 1. Estrazione delle caratteristiche. Per ogni lavoro, viene indicato se dopo l'estrazione delle caratteristiche vengono eseguiti altri compiti. I risultati riportati sono relativi al compito successivo all'estrazione delle caratteristiche. Le abbreviazioni sono utilizzate per risonanza magnetica (RM), lombalgia (LBP), accuratezza (Acc), errore assoluto medio (MAE), apprendimento automatico (ML), macchina vettoriale di supporto (SVM).

I dischi intervertebrali (IVD) sono le strutture lombari più studiate (cinque studi), seguite dalle vertebre (tre studi), mentre uno studio ha valutato la LBP senza concentrarsi su una struttura specifica. Vale la pena notare che solo due degli otto articoli si sono concentrati esclusivamente sull'estrazione delle caratteristiche, ovvero il lavoro di Raudner et al. [21] in cui viene presentato il metodo GRAPPATINI per l'estrazione delle caratteristiche degli IVD dalla risonanza magnetica, e il lavoro di Abdollah et al. [22] in cui vengono usati un Random Forest e un'analisi delle texture sulla risonanza magnetica per l'estrazione rispettivamente delle caratteristiche di IVD e vertebre. I restanti sei articoli hanno descritto l'esecuzione di ulteriori compiti dopo l'estrazione delle caratteristiche. In particolare, quattro di essi eseguono compiti di classificazione, uno di regressione e uno di segmentazione.

Tutti i lavori che hanno eseguito ulteriori compiti dopo l'estrazione delle caratteristiche hanno usato tecniche di machine learning piuttosto che di deep learning: questo è uno dei vantaggi dell'estrazione delle caratteristiche, in quanto permette di ottenere risultati utilizzando metodi molto più veloci e meno dispendiosi dal punto di vista computazionale. Per quanto riguarda la classificazione, Adankon et al. [19] sono stati gli unici a utilizzare immagini 3D della superficie della schiena umana: hanno estratto le caratteristiche di 165 pazienti utilizzando descrittori geometrici locali e le hanno inviate a una *Support Vector Machine* (SVM)

ai minimi quadrati per la classificazione dei tipi di curve scoliotiche, ottenendo un'accuratezza del 95%. Yang et al. [8] hanno utilizzato una trasformata wavelet di Gabor per estrarre caratteristiche dalla risonanza magnetica di 109 soggetti e un tracciatore di caratteristiche Kanade-Lucas-Tomasi (KLT) per identificare le alterazioni degenerative lombari con un'accuratezza dell'88,3%. Ruiz-España et al. [23] hanno estratto caratteristiche dalla risonanza magnetica di 67 pazienti utilizzando il *Gradient Vector Flow* e hanno testato diversi modelli di apprendimento automatico per classificare le IVD degenerate, ottenendo un'accuratezza superiore al 90%. Ketola et al. [24] hanno eseguito l'estrazione di caratteristiche di texture da 518 risonanze magnetiche e hanno utilizzato la regressione logistica per discriminare tra LBP sintomatico e asintomatico con un'accuratezza dell'83%.

Per quanto riguarda il compito di regressione, Garcia-Cano et al. [10] hanno estratto le caratteristiche dalle immagini radiografiche di 150 pazienti attraverso l'analisi delle componenti indipendenti e hanno utilizzato la regressione Random Forest per prevedere la progressione della curva spinale negli adolescenti con scoliosi idiopatica, ottenendo un MAE di 4,79° per l'angolo di Cobb.

Per quanto riguarda il compito di segmentazione, Castro-Mateos et al. [20] hanno estratto le caratteristiche dalla risonanza magnetica 3D di 59 soggetti e hanno eseguito la segmentazione degli IVD utilizzando lo spazio dei modelli statistici di forma e lo spazio B-Spline, ottenendo un punteggio medio DICE dell'88,4%.

4.2 SEGMENTAZIONE

La segmentazione delle immagini è il compito di dividere un'immagine in sottoregioni corrispondenti a diversi elementi dell'immagine, con l'obiettivo di identificare con precisione i confini dei diversi elementi dell'immagine. Questo approccio solitamente sfrutta immagini segmentate manualmente per addestrare un modello di intelligenza artificiale. Diversi manoscritti

inclusi nella revisione hanno svolto un compito di segmentazione e alcuni hanno utilizzato la segmentazione come passo preliminare per ulteriori compiti. Per questo motivo, nelle prossime sezioni riporteremo, ove possibile, non solo i risultati della segmentazione, ma anche quelli dei compiti successivi per i quali la segmentazione viene utilizzata con l'obiettivo di localizzare e/o identificare le strutture. In questa rassegna, ci riferiamo al compito di individuare componenti specifiche (ad esempio, vertebre) come "localizzazione", mentre ci riferiamo al compito di assegnare un'etichetta a componenti specifiche (ad esempio, L1, L2, ecc.) come "identificazione". Inoltre, abbiamo differenziato gli articoli inclusi in base al fatto che sfruttassero tecniche DIP o NN. In questa revisione, abbiamo identificato 38 articoli che utilizzavano tecniche DIP e 23 che utilizzavano NN. Tuttavia, vale la pena notare come la maggior parte degli sforzi di ricerca recenti si stia spostando verso le tecniche di deep learning: prendendo in considerazione gli articoli pubblicati negli ultimi 5 anni (2016-2021), questa revisione include 16 articoli che utilizzano DIP e 23 che utilizzano NN.

4.2.1 Elaborazione digitale delle immagini

Le tecniche di segmentazione DIP elaborano le immagini digitali per trovare i confini delle diverse regioni in base a caratteristiche semantiche, sfruttando metodi come la soglia del gradiente o i modelli statistici di forma. In questa revisione, abbiamo identificato un totale di 38 articoli che hanno eseguito la segmentazione DIP su diversi tipi di immagini (Tabella 2):

- 15 articoli sulla risonanza magnetica (2 dei quali hanno preso in considerazione la risonanza magnetica 3D);
- 15 articoli su immagini TC;
- 1 articoli su immagini di risonanza magnetica e TC;
- 3 articoli sulle immagini fluoroscopiche;
- 2 articoli sulle immagini a ultrasuoni;
- 2 articoli sulle immagini a raggi X.

Autore/ Anno	Compito principale	Tipo di dati	# Pazienti	Strutture coinvolte	Risultati	Modello
Haq, 2015 [26]	Segmentazione	Risonanza magnetica 3D	21	Dischi	DICE = 91,7%	Modelli consapevoli della forma
Neubert, 2012 [27]	Segmentazione e identificazione	Risonanza magnetica 3D	28	Dischi e vertebre	DICE = 89 e 91%, Sen = 100%, Spec = 98%	Modello statistico di forma
Haq, 2020 [25]	Segmentazione	Immagini TC	18 SpineWeb	Dischi	DICE = dal 91,7 al 95,4%.	Modello deformabile con statistiche di forma
Li, 2018 [28]	Segmentazione	Immagini TC	115 (Microsoft R.+ SpineWeb)	Vertebre	DICE = 92,1%	Modello di miscela gaussiana + soglia
Ibragimov, 2017 [29]	Segmentazione	Immagini TC	30 vertebre	Vertebre	DICE = 84,7%	Rilevamento dei punti di riferimento e modelli deformabili
Yu, 2018 [30]	Segmentazione	Immagini TC	21 immagini	Vertebre	DICE = 93,9%	Taglio a griglia assistito da foglio osseo
Korez, 2015 [31]	Segmentazione	Immagini TC	220	Vertebre	DICE = 94,6%	Modello deformabile vincolato alla forma
Al-Helo, 2011 [32]	Segmentazione	Immagini TC	50	Vertebre	Valutazione visiva	Modelli di forma attivi e GVF-snake
Ruiz-España, 2015 [33]	Segmentazione	Immagini TC	10	Vertebre	DICE = 95%	Filtro gaussiano binario selettivo Set di livelli regolarizzati
Huang, 2013 [34]	Segmentazione	Immagini TC	56	Vertebre	DICE = 94%	Soglia di Otsu, set di livelli basati su bordi e regioni
Mahdy, 2018 [35]	Segmentazione e localizzazione	Immagini TC	10	Vertebre	Valutazione visiva	Soglia e K-Means adattivo
Courbot, 2016 [36]	Localizzazione	Immagini TC	15	Vertebre	Valutazione visiva, Acc = 89,4%	Segmentazione a catena di Markov nascosta
Rasoulian, 2013 [37]	Localizzazione	Immagini TC	32	Vertebre	Valutazione visiva, centro di massa MAE = 2 mm	Modello di forma multi-oggetto
Mastmeyer, 2006 [38]	Segmentazione	Immagini TC	41	Vertebre	DICE > 98,6%	Crescita dei volumi e operazioni morfologiche
Jimenez-Pastor, 2020 [39]	Localizzazione e identificazione	Immagini TC	272 immagini	Vertebre	Errore di localizzazione = 13,7 mm, Acc = 74,8%.	Foresta decisionale + elaborazione morfologica delle immagini

Lee, 2011 [40]	Localizzazione e identificazione	Immagini TC	19	Vertebre	Errore di localizzazione = 0,14 mm, Acc = 93,2%.	Costo integrato threshold and thinning-based
Klinder, 2009 [41]	Localizzazione e identificazione	Immagini TC	64	Vertebre	Errore di localizzazione = 1,1 mm, Acc = 92%.	Modelli di forma triangolari
Štern, 2009 [42]	Localizzazione	Immagini di risonanza magnetica e TC	13 e 29 immagini	Dischi e vertebre	Errore di localizzazione = 2,8 e 1,8 mm	Analisi della geometria delle strutture spinali
Wong, 2008 [43]	Segmentazione e tracciamento	Immagini fluoroscopiche	2 video	Vertebre	Valutazione visiva	Wavelet e contorno attivo della forma
Zheng, 2011 [44]	Segmentazione e ricostruzione 3D	Immagini fluoroscopiche	4	Vertebre	Errore medio di ricostruzione<1,6 mm	Modelli statistici di forma
Michopoulou, 2009 [45]	Segmentazione	Risonanza magnetica	34	Dischi	DICE = 90%	Atlas-robust-fuzzy C-Means
Fallah, 2018 [46]	Segmentazione	Risonanza magnetica	50	Dischi e vertebre	DICE = 92,5 e 91,4%	Campo casuale condizionale gerarchico e random forest
Ghosh, 2014 [47]	Segmentazione	Risonanza magnetica	212	Dischi e vertebre	DICE = 87 e 84%	Random forest e caratteristiche del contesto
Kim, 2018 [48]	Segmentazione	Risonanza magnetica	19	Vertebre	DICE = 90%	Algoritmi di segmentazione basati su grafi e linee
Gaonkar, 2017 [49]	Segmentazione	Risonanza magnetica	63	Vertebre	DICE = 83%	Insieme multiparametrico
Gawel, 2018 [50]	Segmentazione	Risonanza magnetica	50	Vertebre	DICE = 91,4%	Classificatore a cascata e Modello di aspetto attivo
Engstrom, 2011 [51]	Segmentazione	Risonanza magnetica	20	Muscoli	DICE = 87%	Modello statistico di forma
Baum, 2018 [52]	Segmentazione	Risonanza magnetica	10	Muscoli	DICE = 83%	Modello di forma media e modello a doppia caratteristica
Zheng, 2004 [53]	Segmentazione	Immagini fluoroscopiche	1	Vertebre	Valutazione visiva	Trasformata di Hough e descrittori di Fourier
Jurcak, 2008 [54]	Segmentazione	Risonanza magnetica	20	Muscoli	DICE = 77%	Probabilistic atlas e profili attivi geodetici

Fortin, 2017 [55]	Segmentazione e regressione	Risonanza magnetica	30	Muscoli	Coefficiente di affidabilità = 97-99%	Soglia
Neubert, 2013 [56]	Segmentazione e localizzazione	Risonanza magnetica	44	Dischi	DADI = 92,3%, AUC = 0,98	Modello di forma attivo, analisi discriminante lineare, SVM
Oktay, 2011 [57]	Localizzazione e identificazione	Risonanza magnetica	40	Dischi	Tasso di localizzazione = 95,4%, Acc = 97%.	Modello probabilistico e SVM
Castro-Mateos, 2016 [58]	Identificazione	Risonanza magnetica	48	Dischi	Sensibilità = 87%	Modello di contorno attivo e Feedforward NN
Kim, 2020 [59]	Localizzazione	Ultrasuoni	50	Muscoli	Differenza di 2 mm	Clustering Fuzzy C-Means
Lui, 2014 [60]	Localizzazione	Ultrasuoni	10	Muscoli	Punteggio F1 = 90,9%	Contorno attivo disaccoppiato
Ribeiro, 2010 [61]	Segmentazione	Raggi X	41	Vertebre	DICE = 91,7%	Filtri Gabor e NN
Sa, 2016 [62]	Localizzazione	Raggi X	30	Vertebre	Tasso di positività reale = 75%	GVF-serpente e SVM

Tabella 2. Segmentazione-Elaborazione digitale delle immagini. Per ogni lavoro viene riportato il compito principale, sia che riguardi solo la segmentazione delle componenti lombari, sia che miri a localizzare parti specifiche (ad esempio, il centro di massa) delle componenti, sia che miri a identificare ogni componente (ad esempio, differenziare le vertebre tra loro). Se vengono analizzate più strutture, i risultati corrispondenti sono riportati nello stesso ordine in cui sono presentate le strutture nella colonna "Strutture coinvolte". Le abbreviazioni sono utilizzate per Risonanza magnetica (MRI), Tomografia computerizzata (CT), Accuratezza (Acc), Sensibilità (Sen), Specificità (Spec), Area sotto la curva (AUC), Gradient Vector Flow (GVF), Support Vector Machine (SVM).

Le vertebre sono le strutture lombari più studiate (26 studi), seguite da IVD (10 studi) e muscoli (6 studi). Vale la pena notare che solo uno [25] dei 21 studi che utilizzano immagini TC, radiografiche o fluoroscopiche non prevedeva la segmentazione delle strutture vertebrali. In totale, 20 articoli si sono concentrati solo sulla segmentazione senza ulteriori compiti. Tra gli altri, 12 hanno eseguito la localizzazione successiva delle strutture, 6 hanno condotto l'identificazione successiva delle strutture (4 dei quali hanno eseguito sia la localizzazione che l'identificazione),

mentre la regressione, il tracciamento e la ricostruzione 3D sono stati studiati rispettivamente da 1 studio cadauno.

Per quanto riguarda gli articoli che si sono concentrati esclusivamente sulla segmentazione, Haq et al. [26] hanno utilizzato modelli shape-aware su RM 3D di 21 pazienti per la segmentazione delle IVD, ottenendo un DICE medio del 91,7%. Inoltre, in un articolo successivo, Haq et al. [25] hanno utilizzato un modello deformabile statistico di forma per la segmentazione di IVD su immagini TC di 18 soggetti del dataset SpineWeb, ottenendo punteggi DICE compresi tra il 91,7 e il 95,4%. Li et al. [28] hanno applicato una soglia ai risultati di un modello a miscela gaussiana per la segmentazione delle vertebre su un totale di 115 immagini TC provenienti dai dataset SpineWeb e Microsoft Research, con un DICE medio del 92,1%. Ibragimov et al. [29] hanno utilizzato il rilevamento di punti di riferimento e modelli deformabili per segmentare 30 vertebre su immagini TC, con un DICE dell'84,7%. Yu et al. [30] hanno utilizzato il taglio a griglia assistito da fogli ossei per segmentare vertebre da 21 immagini TC, ottenendo un DICE medio del 93,9%. Korez et al. [31] hanno applicato un modello deformabile shape-constrained per la segmentazione delle vertebre da immagini TC di 220 pazienti, con un DICE del 94,6%. Al-Helo et al. [32] hanno combinato modelli Active-shape e GVF-snake per la segmentazione delle vertebre da immagini TC di 50 soggetti, valutando la qualità della segmentazione mediante una valutazione visiva. Ruiz-España et al. [33] hanno utilizzato un Selective Binary Gaussian Filtering Regularized Level Set per segmentare le vertebre su immagini TC di 10 soggetti, ottenendo un DICE medio del 95%. Huang et al. [34] hanno sfruttato il thresholding di Otsu e i set di livelli basati su bordi e regioni per segmentare le vertebre su immagini TC di 56 soggetti, con un DICE del 94%. Mastmeyer et al. [38] hanno utilizzato operazioni di volume growing e morfologiche per segmentare le vertebre su immagini TC di 41 soggetti, ottenendo punteggi DICE superiori al 98,6%. Zhang et al. [53] hanno applicato la trasformata di Hough e i descrittori di Fourier per la segmentazione delle vertebre su un'immagine fluo-

roscopica, valutando la qualità della segmentazione mediante una valutazione visiva. Michopoulou et al. [45] hanno utilizzato un Atlas-robust-fuzzy C-Means per la segmentazione di IVD su RM di 34 soggetti, ottenendo un DICE del 90%. Fallah et al. [46] hanno sfruttato campi casuali condizionali gerarchici e una foresta casuale per la segmentazione di IVD e vertebre, rispettivamente, sulla risonanza magnetica di 34 soggetti, ottenendo un DICE del 92,5 e 91,4%, rispettivamente. Ghosh et al. [47] hanno combinato Random Forest e caratteristiche contestuali per la segmentazione di IVD e vertebre, rispettivamente, su RM di 212 soggetti, ottenendo un DICE dell'87 e dell'84%, rispettivamente. Kim et al. [48] hanno utilizzato algoritmi di segmentazione basati su grafi e linee per segmentare le vertebre su RM di 19 pazienti, ottenendo un DICE del 90%. Gaonkar et al. [49] hanno applicato diversi parametri per segmentare le vertebre su RM di 63 soggetti, con un DICE medio dell'83%. Gawel et al. [50] hanno combinato un classificatore a cascata e un modello di Active Appearance per segmentare le vertebre su 50 RM, ottenendo un DICE del 91,4%. Engstrom et al. [51] hanno utilizzato un modello di forma statistica per la segmentazione del muscolo quadratus lumborum sulla risonanza magnetica di 20 pazienti, ottenendo un DICE dell'87%. Baum et al. [52] hanno sfruttato un modello Average Shape e un modello Dual Feature per la segmentazione dei muscoli paraspinali su RM di 10 soggetti, con un DICE dell'83%. Jurcak et al. [54] hanno applicato atlanti probabilistici e contorni attivi geodetici per la segmentazione del muscolo quadrato lombare sulla risonanza magnetica di 20 soggetti, con un DICE del 77%. Ribeiro et al. [61] hanno utilizzato filtri di Gabor e una RNA per segmentare le vertebre su immagini radiografiche di 41 pazienti, ottenendo un DICE del 91,7%.

Per quanto riguarda gli articoli che hanno eseguito la localizzazione dopo la segmentazione, Mahdy et al. [35] hanno utilizzato un metodo a soglia seguito da un K-Means adattivo per la segmentazione e la localizzazione delle vertebre lombari su immagini TC di 10 soggetti, al fine di identificare gli IVD degenerati, e hanno valutato le prestazioni mediante una valu-

tazione visiva. Courbot et al. [36] hanno sfruttato una catena di Markov nascosta per la segmentazione semi-automatica delle vertebre su immagini TC di 15 soggetti, ottenendo un'accuratezza di localizzazione dell'89,4%. Rasoulian et al. [37] hanno sviluppato un modello di forma multi-oggetto per la localizzazione delle vertebre su 32 immagini TC, localizzando correttamente i centri di massa con un MAE di 2 mm allo scopo di identificare la posizione ottimale per l'iniezione dell'ago spinale. Štern et al. [42] hanno eseguito un'analisi della geometria delle strutture spinali per localizzare i centri di IVD e vertebre su 13 immagini MRI e 29 TC, rispettivamente, con un errore di localizzazione di 2,8 e 1,8 mm. Neubert et al. [56] hanno utilizzato un modello Active Shape per segmentare le IVD sulla risonanza magnetica di 44 soggetti, ottenendo un DICE del 92,3% e una AUC di 0,98 per la localizzazione delle IVD degenerate utilizzando l'analisi discriminante lineare e SVM. Kim et al. [59] hanno sfruttato il Fuzzy C-Means Clustering per la localizzazione del muscolo multifido lombare su immagini ecografiche di 50 soggetti, con una discrepanza di localizzazione di 2 mm. Lui et al. [60] hanno utilizzato Decoupled Active Contour per la localizzazione del muscolo multifido lombare su immagini ecografiche di 10 soggetti, ottenendo un F1-Score del 90,9%. Sa et al. [62] hanno utilizzato Gradient Vector Flow Snake e SVM per la localizzazione delle vertebre su immagini radiografiche di 30 soggetti, ottenendo un tasso di veri positivi del 75%.

Per quanto riguarda gli studi che hanno identificato dopo la segmentazione, Neubert et al. [27] hanno utilizzato un modello Statistical Shape su RM 3D di 28 soggetti per segmentare e identificare le IVD e le vertebre, ottenendo un DICE di segmentazione dell'89 e 91%, rispettivamente, e una specificità del 98,3% e una sensibilità del 100% per l'identificazione delle IVD degenerate. Castro-Mateos et al. [58] hanno descritto un Active Contour Model per la segmentazione e un Feedforward NN per l'identificazione e la classificazione di IVD su RM di 48 soggetti, ottenendo una sensibilità dell'87%.

Per quanto riguarda i lavori che hanno eseguito sia la localizzazione che l'identificazione, Jimenez-Pastor et al. [39] hanno utilizzato una foresta decisionale e l'elaborazione morfologica delle immagini per localizzare e identificare le vertebre su 272 immagini TC, ottenendo un errore di localizzazione di 13,7 mm e un'accuratezza del 74,8%. Lee et al. [40] hanno sfruttato la soglia e il costo integrato basato sull'assottigliamento su immagini TC di 19 soggetti, per la localizzazione e l'identificazione dei peduncoli lombari al fine di aumentare l'accuratezza e la sicurezza durante il posizionamento delle viti transpedicolari, con un errore di localizzazione di 0,14 mm e un'accuratezza del 93,2%. Klinder et al. [41] hanno utilizzato un modello Triangulated Shape su immagini TC di 64 soggetti, ottenendo un errore di localizzazione delle vertebre di 1,1 mm e un'accuratezza del 92%. Oktay et al. [57] hanno combinato un modello probabilistico con un SVM per localizzare e rilevare gli IVD sulla risonanza magnetica di 40 soggetti, ottenendo un tasso di localizzazione del 95,4% e un'accuratezza del 97%.

Inoltre, Wong et al. [43] hanno utilizzato Wavelets e un modello Shape-Active Contour-Based per la segmentazione e il tracciamento delle vertebre su 2 video di immagini fluoroscopiche, valutando le prestazioni visivamente. Zheng et al. [44] hanno utilizzato modelli di forma statistica per la segmentazione delle vertebre e la ricostruzione 3D su 4 immagini fluoroscopiche, ottenendo un errore medio di ricostruzione inferiore a 1,6 mm. Infine, Fortin et al. [55] hanno utilizzato un algoritmo di sogliatura per la segmentazione e la quantificazione della composizione dei muscoli paraspinali con un coefficiente di affidabilità compreso tra il 97 e il 99%.

4.2.2 Deep Learning

Il deep learning è una classe di algoritmi di intelligenza artificiale basati sulle reti neurali. Più in dettaglio, una NN si dice "profonda" se è composta da più di 2 strati nascosti. Le tecniche di apprendimento profondo per la segmentazione prendono

come input l'intera immagine originale ed eseguono l'estrazione delle caratteristiche, la selezione delle caratteristiche, la segmentazione e qualsiasi altra fase (ad esempio, classificazione, regressione) in un unico modello. In questa rassegna, abbiamo identificato un totale di 23 lavori che hanno eseguito la segmentazione tramite deep learning, le cui caratteristiche principali sono riportate nella Tabella 3. In dettaglio:

- 13 articoli sulla risonanza magnetica (2 dei quali hanno preso in considerazione la risonanza magnetica 3D e 1 con l'aggiunta di note cliniche);
- 5 articoli su immagini TC;
- 4 articoli sulle immagini a raggi X (di cui 1 in combinazione con le immagini Moire);
- 1 articolo sulle immagini a ultrasuoni.

Autore/ Anno	Compito principale	Tipo di dati	# Pazienti	Strutture coinvolte	Risultati	Modello
Iriondo, 2020 [63]	Segmentazione	Risonanza magnetica 3D	31	Dischi	DICE > 85%	Memoria di contesto da grossolana a fine NN
Staartjes, 2021 [64]	Segmentazione e ricostruzione	Risonanza magnetica 3D	3	Tutte le strutture	Valutazione visiva	CNN
Lee, 2020 [65]	Segmentazione e ricostruzione	Immagini TC	280 immagini	Tutte le strutture	MAE = 21 pixel	Reti avversarie generative
Fan, 2020 [66]	Segmentazione e ricostruzione	Immagini TC	108	Tutte le strutture	Triangolo Kambin = 161 mm2	Rete U
Malinda, 2020[67]	Segmentazione	Immagini TC	120	Vertebre	DICE = 94,2%	Reti generative avversarie
Siemionow, 2020 [68]	Identificazione	Immagini TC	45	Vertebre	Acc = 96-99%	CNN
Netherton, 2020 [69]	Localizzazione e identificazione	Immagini TC	330 immagini	Vertebre	Errore di localizzazione = 2,2 mm, Acc = 94%.	Insieme di reti X
Watanabe 2019 [70]	Regressione	Immagini moire + raggi X	1996	Vertebre	Angolo di Cobb MAE = 3,42°	CNN
Kim, 2018 [71]	Segmentazione	Risonanza magnetica	SpineWeb 20	Dischi	DICE = 89,4%	CNN (BSU-net)

Shen, 2021 [72]	Segmentazione	Risonanza magnetica	120	Dischi, canale vertebrale e muscoli	Jaccard: 87, 82 e 85%	Feedforward NN
Gaonkar, 2019 [73]	Segmentazione	Risonanza magnetica	39295	Dischi e canale vertebrale	DICE = 88 e 87%	Dischi: U-net, Canale: SVM e RT
Huang, 2020 [74]	Segmentazione	Risonanza magnetica	100	Dischi e vertebre	Jaccard = 92,6 e 94,7%	Rete U
Li, 2021 [75]	Segmentazione	Risonanza magnetica	120	Vertebre e canale spinale	DICE = 92,5%	CNN
Li, 2019 [76]	Segmentazione	Risonanza magnetica	120	Muscoli	DICE > 91,3%	Rete a U deformata
Zhou, 2020 [77]	Segmentazione	Risonanza magnetica	57	Vertebre	DICE = 84,9%	Rete U
Jamaludin, 2017 [78]	Classificazione	Risonanza magnetica	2009	Dischi e vertebre	Acc = 95,6%	CNN
Natalia, 2020 [79]	Regressione	Risonanza magnetica	515	Dischi e canale vertebrale	Errore medio: 0,9 mm	SegNet e algoritmo Contour Evolution
Zhou, 2019 [80]	Identificazione	Risonanza magnetica	1318	Vertebre	Acc = 98,9%	CNN
Forsberg, 2017 [81]	Identificazione	Risonanza magnetica con note cliniche	475	Vertebre	Acc = 97%	CNN e modelli grafici basati sulle parti
Baka, 2017 [82]	Identificazione	Ultrasuoni	19 set di dati	Vertebre	Acc = 92%	CNN e strategia di abbinamento
Cho, 2020 [83]	Segmentazione e regressione	Raggi X	629	Vertebre	DADI = 82,1%, MAE = 8.055°	Rete U
Li, 2016 [84]	Identificazione	Raggi X	110	Vertebre	Acc = 80,4%	CNN
Sa, 2017 [85]	Localizzazione	Raggi X	1081 immagini	Dischi	Precisione = 90,5%	R-CNN più veloce

Tabella 3. Segmentazione-Apprendimento profondo. Per ogni lavoro è riportato il compito principale, che sia la segmentazione delle componenti lombari, oppure la localizzazione di parti specifiche (ad esempio, il centro di massa) delle componenti, o l'identificazione di ogni componente (ad esempio, differenziare le vertebre tra loro). Se vengono analizzate più strutture, i risultati corrispondenti sono riportati seguendo lo stesso ordine con cui le strutture sono presentate nella colonna "Strutture coinvolte". Le abbreviazioni sono utilizzate per risonanza magnetica (MRI), tomografia computerizzata (CT), errore assoluto medio (MAE), accuratezza (Acc), rete neurale convoluzionale (CNN), macchina vettoriale di supporto (SVM), alberi di regressione (RT).

Le vertebre sono state le strutture lombari più studiate (16 articoli), seguite da IVD (11 articoli), canale spinale (7 articoli) e muscoli (5 articoli). In totale, 9 articoli si sono concentrati esclusivamente sulla segmentazione senza ulteriori compiti. Tra gli altri, 5 manoscritti hanno eseguito l'identificazione di strutture successive, 3 hanno eseguito un compito di regressione, 3 hanno eseguito la ricostruzione di strutture successive, 1 lavoro ha eseguito la classificazione, 1 ha eseguito la localizzazione di strutture e 1 ha eseguito sia la localizzazione che l'identificazione di strutture. Vale la pena notare che la stragrande maggioranza dei lavori inclusi in questa sezione ha sfruttato le reti neurali convoluzionali (CNN) o i modelli che ne derivano.

Per quanto riguarda gli articoli che si sono concentrati esclusivamente sulla segmentazione, Iriondo et al. [63] hanno utilizzato una NN a memoria di contesto da grossolana a fine per segmentare gli IVD su RM 3D di 31 soggetti, ottenendo una DICE superiore all'85%. Malinda et al. [67] hanno utilizzato reti avversarie generative (GAN) per la segmentazione delle vertebre su immagini TC di 120 soggetti, ottenendo un DICE del 94,2%. Kim et al. [71] hanno sfruttato una rete BSU per la segmentazione delle vertebre su 20 risonanze magnetiche del dataset SpineWeb, ottenendo un DICE dell'89,4%. Shen et al. [72] hanno utilizzato una Feedforward NN su RM di 120 soggetti, ottenendo un indice di Jaccard per la segmentazione di IVD, canale spinale e muscoli rispettivamente dell'87, 82 e 85%. Gaonkar et al. [73] hanno applicato una rete a U per segmentare i DPI su 39295 immagini di risonanza magnetica, ottenendo un DICE dell'88%; hanno inoltre combinato una SVM con un albero di regressione per segmentare il canale spinale con un DICE dell'87%. Huang et al. [74] hanno utilizzato una rete a U per segmentare IVD e vertebre su 100 risonanze magnetiche, ottenendo un indice di Jaccard rispettivamente del 92,6 e del 94,7%. Li et al. [75] hanno utilizzato una CNN per segmentare vertebre e canale spinale sulla risonanza magnetica di 120 pazienti, ottenendo un DICE complessivo del 92,5%. Inoltre, hanno utilizzato una rete a U deformata [76] per la segmentazione dei muscoli paraspinali su 120 RM ottenendo

un DICE complessivo superiore al 91,3%. Zhou et al. [77] hanno utilizzato una rete a U per la segmentazione delle vertebre su RM di 57 soggetti, ottenendo un DICE dell'84,9%.

Per quanto riguarda i lavori che hanno eseguito l'identificazione della struttura dopo la segmentazione, Siemionow et al. [68] hanno utilizzato una CNN per identificare le vertebre su immagini TC di 45 soggetti, con un'accuratezza complessiva compresa tra il 96 e il 99%. Zhou et al. [80] hanno combinato una CNN e la somiglianza con un'immagine lombare precedente per l'identificazione delle vertebre su immagini MRI di 1318 soggetti sani e non, ottenendo un'accuratezza del 98,9%. Forsberg et al. [81] hanno combinato una CNN e modelli grafici basati su grafi su immagini di risonanza magnetica arricchite con note cliniche per identificare le vertebre di 475 pazienti, ottenendo un'accuratezza del 97%. Baka et al. [82] hanno utilizzato una CNN e una strategia di matching per l'identificazione delle vertebre su immagini ecografiche provenienti da 19 set di dati, ottenendo un'accuratezza del 92%. Li et al. [84] sono stati gli unici a eseguire l'identificazione delle vertebre su immagini radiografiche. Hanno applicato una CNN su 110 immagini, ottenendo un'accuratezza dell'80,4%.

Per quanto riguarda gli articoli che hanno eseguito un compito di regressione, Watanabe et al. [70] hanno utilizzato una CNN per stimare l'allineamento spinale su immagini Moire del 1996, con un MAE dell'angolo di Cobb di 3,42°. Natalia et al. [79] hanno combinato una SegNet e un algoritmo di evoluzione dei contorni per misurare il diametro anteroposteriore e l'ampiezza foraminale su RM di 515 pazienti affetti da stenosi spinale lombare, con un errore medio di 0,9 mm. Cho et al. [83] hanno utilizzato una rete a U per la segmentazione e la misurazione automatizzata della lordosi lombare su immagini radiografiche di 629 pazienti, ottenendo un DICE dell'82,1% e un MAE di 8,06°.

Per quanto riguarda gli articoli che eseguono un compito di ricostruzione, Staartjes et al. [64] hanno sviluppato una CNN per segmentare e ricostruire le strutture lombari dalla risonanza magnetica 3D di 3 pazienti, valutando le prestazioni mediante

una valutazione visiva. Lee et al. [65] hanno utilizzato GAN per generare strutture lombari sintetiche di risonanza magnetica della colonna vertebrale da 280 immagini TC, con un MAE di 21 pixel. Fan et al. [66] hanno utilizzato una rete a U per ricostruire le strutture lombari da immagini TC di 108 soggetti, con un triangolo di Kambin di 161 mm.

Per quanto riguarda gli articoli che svolgono un compito di classificazione, Jamuladin et al. [78] hanno utilizzato una CNN per la classificazione di IVD e vertebre sulla risonanza magnetica di 2009 soggetti, ottenendo un'accuratezza del 95,6%.

Inoltre, Sa et al. [85] hanno messo a punto una Faster Region-based CNN (R-CNN) per la localizzazione di IVD su 1081 immagini radiografiche con una precisione del 90,5%. Infine, Netherton et al. [69] hanno utilizzato un insieme di reti X per localizzare e identificare le vertebre su 330 immagini TC, ottenendo un errore di localizzazione di 2,2 mm e una precisione del 94%.

5. DISCUSSIONE

A causa dell'ampio uso di tecniche avanzate di imaging e della complessità delle strutture anatomiche coinvolte nello sviluppo del LBP e delle sue sequele, la ricerca ha molto studiato l'utilizzo dell'IA nell'elaborazione di immagini digitali per diversi scopi. La stragrande maggioranza degli studi usa la risonanza magnetica o la TAC, mentre una minoranza usa le immagini radiografiche, fluoroscopiche o ecografiche. Vale la pena notare che le strutture vertebrali sono l'oggetto principale degli articoli che eseguono la segmentazione, sia con tecniche DIP sia con tecniche di deep learning; viceversa, gli articoli che eseguono l'estrazione di caratteristiche si concentrano principalmente sugli IVD.

Per quanto riguarda l'estrazione delle caratteristiche, ovvero la capacità di un sistema di riconoscere un insieme specifico di caratteristiche rilevanti, tutti gli studi inclusi hanno mostrato complessivamente un'accuratezza superiore all'80% nell'identificare la posizione delle vertebre [24] e degli IVD [8,20,21,22,23,24],

con la capacità di rilevare anche le lacerazioni dell'anulus e l'ernia del disco lombare [21,22,23]. Sebbene la maggior parte degli studi sia stata condotta su immagini di risonanza magnetica [8,20,21,22,23,24], uno studio ha utilizzato immagini a raggi X [10] e un altro ha costruito un modello 3D della schiena dei pazienti utilizzando una tecnologia di acquisizione di superfici non invasiva [19]. Inoltre, alcuni di questi studi hanno anche riportato la capacità dei sistemi descritti di eseguire compiti di classificazione e regressione sui dati estratti, come la stima del grado di degenerazione dell'IVD [8,22,23,24], la classificazione del tipo di curva scoliosica [19] e la previsione della progressione della curva [10], la presenza di stenosi spinale [23] e l'esplorazione della correlazione tra i cambiamenti degenerativi e la presenza di LBP [24].

Tuttavia, la maggior parte degli studi si è concentrata sulla segmentazione, ovvero sulla differenziazione di specifiche sottoregioni di un'immagine in base a parametri distinti. Tradizionalmente, i compiti di segmentazione sono stati eseguiti dai sistemi DIP attraverso la suddivisione degli elementi all'interno di un'immagine basata sulla soglia del gradiente o su modelli statistici di forma, che rientrano nella definizione di segmentazione semantica [86]. Tuttavia, la ricerca recente ha esplorato l'uso di sistemi di intelligenza artificiale basati sull'apprendimento profondo che sono in grado di eseguire più compiti a livello base e avanzato in un unico modello [1]. Le vertebre sono di gran lunga la struttura più studiata, con sistemi di IA che hanno raggiunto un DICE > 90% e un'accuratezza > 90% nella maggior parte degli studi inclusi nella nostra revisione, sia utilizzando modelli DIP [28,29,30,31,32,33,34,35,36,37,38,39,40,41,43,44,48,49,50,53,61,62] che modelli di deep learning [67,69,77,80,82,83,84]. In particolare, uno studio di Lee et al. [40] ha proposto un modello per ottenere una segmentazione automatizzata dei peduncoli lombari da immagini TC al fine di aumentare l'accuratezza e la sicurezza durante il posizionamento delle viti transpediculari. D'altra parte, uno studio di Watanabe e colleghi [70] ha descritto una CNN in grado di stimare l'allineamento spinale, la rotazione vertebrale e l'angolo di Cobb con un errore assoluto medio di 3,6 pixel per la posizione

vertebrale, di 2,9° per la rotazione vertebrale e di 3,42° per quanto riguarda l'angolo di Cobb stimato. Analogamente, Cho et al. [83] hanno presentato una CNN in grado di segmentare le vertebre lombari e di calcolare successivamente la lordosi lombare, con un errore assoluto medio di 8,055°. In questo studio sono stati descritti anche diversi sistemi di IA per la segmentazione automatizzata degli IVD [25,26,45,48,56,57,58,63] con un DICE > 90% in quasi tutti gli studi. Inoltre, le prestazioni dei sistemi sviluppati per la segmentazione dei muscoli paraspinali hanno riportato una maggiore variabilità rispetto ad altre strutture [51,52,54,55,60,71], con valori DICE più elevati per i sistemi basati su modelli di deep learning [76]. Inoltre, alcuni studi hanno valutato la segmentazione simultanea di più strutture, in particolare IVD e vertebre [27,42,46,56,74,78], con un DICE > 90% nei sistemi basati su DIP [27,42,46,56] e un'accuratezza > 95% nella maggior parte dei sistemi basati su deep learning [68,71,74,78,85]. Inoltre, alcuni di questi ultimi sono stati utilizzati per sintetizzare immagini TC da RM e viceversa. Ad esempio, Staartjes et al. [64] hanno introdotto un sistema basato su CNN in grado di generare immagini TC sintetiche dalla risonanza magnetica della colonna vertebrale, in modo da acquisire informazioni più precise sulle strutture ossee rispetto alla risonanza magnetica tradizionale senza dover esporre i pazienti a radiazioni aggiuntive. Lee e colleghi [65] hanno invece presentato un modello basato su GAN in grado di produrre una risonanza magnetica sintetica a partire da scansioni TC della colonna vertebrale, ottenendo una somiglianza complessiva media con le scansioni RM reali dell'80,2%. Questo studio ha dimostrato la possibilità di estrarre informazioni accurate sui tessuti molli dalla TAC della colonna vertebrale senza la necessità di prenotare una risonanza magnetica, che spesso è costosa e richiede molto tempo. Altri studi hanno dimostrato la possibilità di calcolare automaticamente l'area del canale spinale [73] e di segmentare e ricostruire più strutture contemporaneamente [47,66,72,75,79] con un elevato grado di accuratezza.

La Figura 4 mostra un boxplot che riassume i risultati per la segmentazione di IVD, vertebre e muscoli lombari e l'accuratezza

dell'identificazione per le diverse strutture lombari. Per quanto riguarda la segmentazione di IVD e vertebre, vale la pena notare che le tecniche DIP e deep learning ottengono risultati molto simili, con i metodi DIP leggermente migliori. Ciò è dovuto principalmente alla superficie regolare e omogenea di queste strutture, i cui bordi ben definiti possono essere identificati efficacemente con tecniche DIP, come i metodi a soglia e di region-growing. Al contrario, le prestazioni di segmentazione del muscolo lombare delle tecniche di deep learning sono sensibilmente migliori di quelle dei metodi DIP. In effetti, la struttura dei muscoli è irregolare e più difficile da individuare correttamente, e le NN profonde forniscono uno strumento migliore per questo compito. Per quanto riguarda l'accuratezza dell'identificazione, il deep learning fornisce generalmente risultati migliori; tuttavia, i metodi DIP seguiti da tecniche di apprendimento automatico sono tipicamente più veloci e meno costosi dal punto di vista computazionale e, in alcuni casi, forniscono prestazioni simili.

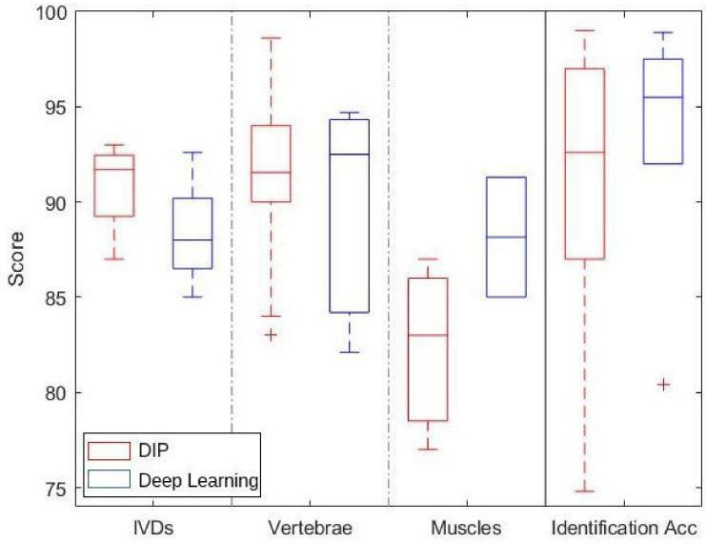


Figura 4. *Boxplot che riassume i risultati per diverse strutture e compiti. Le tre colonne di sinistra si riferiscono ai punteggi DICE per la segmentazione di IVD, vertebre e muscoli; la colonna di destra si riferisce all'accuratezza dell'identificazione per le diverse strutture.*

Sebbene l'applicazione della computer vision all'elaborazione di immagini radiologiche della colonna vertebrale sia in continua crescita, ci sono alcuni dubbi in merito. Infatti, la validazione del sistema dipende ancora in larga misura da diversi interventi dell'utente e non può sostituire la controparte umana per ovvie ragioni, sia dal punto di vista clinico che etico. Inoltre, i metodi più performanti si basano sull'applicazione di NN, che di solito richiedono una grande quantità di immagini e capacità di calcolo per l'addestramento, che non sono disponibili per tutti i ricercatori. Tuttavia, alcune tecniche DIP forniscono prestazioni uguali o migliori nella segmentazione di strutture di forma regolare, come vertebre e IVD, richiedendo al contempo una quantità minore di dati per l'addestramento e limitando il carico computazionale. Inoltre, esistono già alcuni metodi per il rilevamento e la classificazione automatica di problematiche quali spondilolistesi, ernia del disco e scoliosi.

6. CONCLUSIONI

Nell'ultimo decennio, l'utilizzo dell'IA è aumentato notevolmente in tutti i campi e la ricerca medica non ha fatto eccezione. In effetti, i computer basati sull'IA hanno già mostrato il potenziale per rivoluzionare il campo medico, compresa la chirurgia spinale. In questo studio abbiamo esaminato la letteratura disponibile sull'uso dell'IA, e più specificamente della computer vision, nella prevenzione, nella diagnosi e nel trattamento del LBP. In conclusione, le tecniche di computer vision promettono di migliorare efficacemente la pratica clinica negli anni a venire, grazie alla disponibilità di set di dati pubblici e al naturale prossimo aumento della capacità computazionale. Inoltre, si stanno compiendo passi avanti verso l'interpretabilità dell'IA e, in particolare, dei modelli di deep learning. Tali miglioramenti porteranno allo sviluppo di sistemi che non richiederanno i diversi interventi da parte dell'utente, fornendo così un valido strumento di valutazione per i medici. La diagnosi e il trattamento della

LBP richiedono spesso l'utilizzo e l'integrazione di modalità di imaging avanzate. Inoltre, diverse alterazioni strutturali, spesso sottili e non univoche da interpretare, concorrono a definire lo scenario clinico. In questo quadro, l'uso dell'intelligenza artificiale e della computer vision può assistere e implementare efficacemente il processo diagnostico, migliorando così i risultati clinici e l'accuratezza diagnostica.

10 - MACHINE LEARNING ED ELABORAZIONE DEL LINGUAGGIO NATURALE NELLA SALUTE MENTALE: UNA REVISIONE SISTEMATICA

Tratto e tradotto da
Le Glaz A, Haralambous Y, Kim-Dufor D, Lenca P, Billot R, Ryan TC, Marsh J, DeVylder J, Walter M, Berrouiguet S, Lemey C, *Machine Learning and Natural Language Processing in Mental Health: Systematic Review*, J Med Internet Res 2021;23(5):e15708.

doi: 10.2196/15708

Le parti omesse dal curatore rispetto all'originale sono indicate dal segno **[...]**

ABSTRACT

Premessa: i sistemi di machine learning [o "apprendimento automatico"] fanno parte del campo dell'intelligenza artificiale e apprendono automaticamente modelli dai dati per prendere decisioni migliori. L'elaborazione del linguaggio naturale (NLP, [*Natural Language Processing*]), utilizzando corpora e approcci di apprendimento, fornisce buone prestazioni in compiti statistici, come la classificazione dei testi o il sentiment mining.

Obiettivo: l'obiettivo primario di questa revisione sistematica è stato quello di riassumere e caratterizzare, in termini metodologici e tecnici, gli studi che hanno utilizzato tecniche di apprendimento automatico e PNL per la salute mentale. L'obiettivo secondario era quello di considerare il potenziale uso di questi metodi nella pratica clinica della salute mentale.

Metodi: questa revisione sistematica segue le linee guida PRISMA (Preferred Reporting Items for Systematic Review and

Meta-analysis) ed è registrata presso PROSPERO (Prospective Register of Systematic Reviews; numero CRD42019107376). La ricerca è stata condotta utilizzando 4 database medici (PubMed, Scopus, ScienceDirect e PsycINFO) con le seguenti parole chiave: apprendimento automatico, data mining, psichiatria, salute mentale e disturbo mentale. I criteri di esclusione erano i seguenti: lingue diverse dall'inglese, processo di anonimizzazione, studi di casi, articoli di conferenze e revisioni. Non sono state imposte limitazioni sulle date di pubblicazione.

Risultati: sono stati identificati 327 articoli, di cui 269 (82,3%) sono stati esclusi e 58 (17,7%) sono stati inclusi nella revisione. I risultati sono stati organizzati secondo una prospettiva qualitativa. Sebbene gli studi avessero argomenti e metodi eterogenei, sono emersi alcuni temi. Gli studi sulla popolazione potevano essere raggruppati in 3 categorie: pazienti inclusi nei database medici, pazienti che sono andati al pronto soccorso e utenti dei social media. Gli obiettivi principali erano estrarre i sintomi, classificare la gravità della malattia, confrontare l'efficacia della terapia, fornire indizi psicopatologici e mettere in discussione la nosografia attuale. Le cartelle cliniche e i social media sono state le due principali fonti di dati. Per quanto riguarda i metodi utilizzati, la pre-elaborazione ha utilizzato i metodi standard di NLP e l'estrazione di identificatori unici dedicati ai testi medici. Sono stati preferiti classificatori efficienti piuttosto che classificatori dal funzionamento trasparente. Python è stata la piattaforma più utilizzata.

Conclusioni: negli ultimi anni, l'apprendimento automatico e i modelli NLP sono stati temi di grande attualità in medicina e possono essere considerati un nuovo paradigma nella ricerca medica. Tuttavia, questi processi tendono a confermare le ipotesi cliniche piuttosto che a sviluppare informazioni completamente nuove, e solo una categoria importante della popolazione (cioè gli utenti dei social media) è una coorte imprecisa. Inoltre, alcune caratteristiche specifiche della lingua possono migliorare le prestazioni dei metodi NLP e la loro estensione ad altre lingue dovrebbe essere studiata più da vicino. Tuttavia, le tecniche di apprendimento automatico e di PNL forniscono informazioni

utili da dati inesplorati (ad esempio, le abitudini quotidiane dei pazienti, solitamente inaccessibili agli operatori sanitari). Prima di considerare l'IT come un ulteriore strumento di cura della salute mentale, rimangono questioni etiche che dovrebbero essere discusse in fretta. I metodi di machine learning e NLP possono offrire molteplici prospettive nella ricerca sulla salute mentale, ma dovrebbero essere considerati anche come strumenti di supporto alla pratica clinica.

1. INTRODUZIONE

1.1 APPRENDIMENTO AUTOMATICO

I sistemi di machine learning (ML) apprendono automaticamente modelli dai dati per prendere decisioni migliori. In quanto tali, fanno parte di un importante sottocampo dell'intelligenza artificiale (AI). Esistono 3 approcci principali all'apprendimento dai dati: apprendimento supervisionato, non supervisionato e per rinforzo. Nell'apprendimento supervisionato, viene previsto un attributo target e gli algoritmi di ML deducono un modello da dati di input etichettati (cioè un set di dati di addestramento che fornisce esempi descritti da attributi predittivi e valori per l'attributo target). L'obiettivo è fare previsioni su nuovi dati per ottenere buone prestazioni di generalizzazione. Al contrario, nell'apprendimento non supervisionato non esiste un attributo target e quindi non ci sono dati etichettati. L'apprendimento non supervisionato consiste nel dedurre un modello per descrivere modelli nascosti da dati non etichettati. In circostanze in cui l'acquisizione di dati etichettati si rivela difficile (ad esempio, costosa), i metodi di ML semisupervisionati possono utilizzare sia dati etichettati che non etichettati per l'apprendimento. La terza categoria principale di ML è l'apprendimento per rinforzo, in cui il modello ML utilizza un feedback che agisce come premio o punizione per massimizzare le sue prestazioni.

Il metodo ML è limitato a certe capacità. Ad esempio, si basa su raccolte di dati che possono essere incomplete, rumorose o soggette a distorsioni sistematiche, tutte cose che possono portare a previsioni errate. Inoltre, gli algoritmi di ML possono introdurre distorsioni. Le questioni interessanti da affrontare nel ML sono discusse in un articolo di Domingos [1]. Tuttavia, se condotto con attenzione, il ML può avere una grande utilità.

L'IA e il ML hanno numerose applicazioni, molte delle quali si incontrano nella vita quotidiana. Il ML supervisionato, ad esempio, è ampiamente utilizzato per il filtraggio dello spam (cioè per classificare le e-mail in arrivo come spam o non spam) [2]. Viene anche utilizzato per classificare i richiedenti credito in base alle loro probabilità di insolvenza [3]. Il ML non supervisionato, come il clustering degli algoritmi, è in grado di raggruppare i clienti con caratteristiche simili e simile probabilità di acquisto. Questo metodo è ampiamente utilizzato dalle banche per la segmentazione del mercato [4]. Infine, il clustering automatico dei documenti, che organizza documenti simili in classi (per migliorare il reperimento delle informazioni, ad esempio), sta acquisendo importanza a causa del numero crescente di documenti su Internet [5].

Anche l'applicazione del ML nel campo della salute è un problema. Infatti, il ML è ampiamente utilizzato in modelli di malattie critiche nella ricerca cardiologica, neurologica e diabetologica [6] per identificare automaticamente i fattori di rischio delle malattie cardiache [7], per classificare i sottotipi di afasia primaria progressiva [8], e per la caratterizzazione e la diagnosi dei disturbi cognitivi [9], del diabete e dei disturbi cardiovascolari [10-17].

Il ML sta sfidando anche il tradizionale approccio epidemiologico della medicina basata sull'evidenza, grazie alla sua elevata velocità di elaborazione e alla capacità di gestire grandi volumi di dati con variabili eterogenee (cartelle cliniche elettroniche, set di dati amministrativi, sensori indossabili, banche dati genomiche e proteomiche e social media). [18]. In effetti, l'IA e il ML hanno un enorme potenziale per costruire inferenze e trovare modelli in vasti volumi di storie di pazienti, immagini mediche, statisti-

che epidemiologiche e altri particolari come i dati del linguaggio naturale. Ad esempio, possono aiutare i medici a migliorare le diagnosi, a prevedere l'insorgere di malattie e a personalizzare i trattamenti [19,20], a fornire una migliore assistenza ai pazienti [21] e a prevedere l'attività di splicing di singoli esoni e i segni della cromatina dalle sequenze di DNA [22]. Dal punto di vista della salute mentale, la prevenzione del rischio di suicidio è stata recentemente oggetto di studi approfonditi [23-26].

In effetti, anche la cura della salute mentale sta beneficiando dei progressi del ML [27-29]. Il ML classico con dati misti (osservazioni descritte da un misto di variabili numeriche e categoriche) è ampiamente utilizzato, ma i deficit linguistici sono sintomi comuni di depressione, disturbo bipolare, disturbo dello spettro autistico (ASD), disturbo di personalità e schizofrenia [30]. Ciò implica che la linguistica computazionale potrebbe avere un ruolo importante per aiutarci a conoscere la salute mentale e le emozioni degli individui.

Il linguaggio, sia nelle forme parlate che scritte, svolge un ruolo importante nelle applicazioni di ML per la salute mentale. È quindi essenziale capire cosa sia l'elaborazione del linguaggio naturale (NLP) prima di discutere le applicazioni congiunte di ML e NLP nella salute mentale.

1.2 NLP

Il NLP è una sottodisciplina dell'informatica emersa negli anni Sessanta. Nel 1967, il primo libro pubblicato sull'argomento, *Introduction to Computational Linguistics* [31], considera chiaramente il linguaggio da un punto di vista simbolico: descrive tecniche come il parsing della sintassi utilizzando gli alberi di dipendenza o le grammatiche trasformazionali di Chomsky; accenna solo i metodi statistici (come il conteggio delle parole). All'epoca, le risorse di calcolo erano scarse e dovevano essere gestite con attenzione; per questo motivo, un intero capitolo del libro è dedicato all'archiviazione delle grammatiche in memoria. La situazione è cambiata negli anni '90, quando i personal com-

puter sono diventati ampiamente disponibili e sempre più potenti. È emerso un nuovo approccio alla NLP basato su metodi statistici. Il libro di Manning e Schütze, *Foundations of Statistical Natural Language Process* [32], è una pietra miliare di questa evoluzione [32]. Le tre sezioni principali del libro sono dedicate a (1) metodi a livello di parola (collocazioni, *n-grammi* e disambiguazione del senso delle parole), (2) metodi a livello di frase (parsing morfosintattico con modelli di Markov e grammatiche libere dal contesto e probabilistiche) e (3) clustering, classificazione e recupero di informazioni. Le grammatiche probabilistiche libere dal contesto sono un tipico esempio dell'evoluzione dei metodi di PNL: l'approccio simbolico di Chomsky - o almeno una sua versione semplificata - è dotato di probabilità legate alle produzioni, e l'ambiguità del linguaggio naturale si riflette nella coesistenza di diversi alberi sintattici con probabilità diverse.

Nello stesso periodo si sono evoluti anche i metodi simbolici. Gli anni '90 hanno visto l'emergere del World Wide Web, del Semantic Web e dell'ingegneria ontologica. Inizialmente, le due direzioni di ricerca sembravano contraddittorie. La rappresentazione della conoscenza mirava a strutturare la conoscenza in modo simbolico ed esaustivo, mentre il punto di vista statistico considerava il linguaggio nello stesso modo in cui la fisica considera i fenomeni naturali: analizzandoli attraverso vari metodi eterogenei, identificando leggi generali attraverso indicatori numerici e dimostrandole con metodi statistici. Un esempio che illustra quest'ultimo aspetto è l'ipotesi semantica distributiva (originariamente enunciata nell'articolo di Harris intitolato Struttura distributiva [33]), secondo la quale "le parole che ricorrono negli stessi contesti tenderanno ad avere significati correlati". Secondo questa ipotesi, non è necessario identificare il significato preciso di una parola, come richiederebbe un metodo simbolico, ma è sufficiente trovare le cooccorrenze della parola in un corpus e considerarle come semantica della parola. Un metodo molto popolare, chiamato analisi semantica latente (LSA), si basa su quanto segue: la matrice delle occorrenze delle parole nei documenti (contesti) viene ridotta in modo che le dimensio-

ni della nuova matrice rappresentino aggregati di parole e aggregati di documenti, dove ogni dimensione non è interpretabile di per sé, ma quando le parole o i documenti sono rappresentati come vettori in questo nuovo sistema *latente* di coordinate, il prodotto scalare dei vettori può essere utilizzato come misura di correlazione semantica [34]. LSA è anche un esempio di un tipico metodo ML, con una fase di apprendimento (in cui si contano le frequenze delle parole nel corpus e si riduce la matrice delle parole o dei documenti) per eseguire un compito specifico (valutare la somiglianza tra i documenti).

A partire dagli anni 2000 e 2010, si è verificata una nuova evoluzione nell'ambito dell'NLP con l'emergere di reti neurali convoluzionali, ricorrenti e ricorsive (NN) [35]. Utilizzando grandi corpora e approcci di apprendimento sofisticati, questi metodi forniscono buone prestazioni in compiti di natura statistica, come la classificazione dei testi o il sentiment mining. Negli ultimi 3 anni sono stati utilizzati molto di più per l'apprendimento di strutture sintattiche o semantiche superiori (grafi sintattici o concept mining, rispettivamente).

In futuro potrebbero essere utilizzati più frequentemente dei metodi ibridi, che combinano approcci simbolici e statistici. La presenza di metodi ML nei sistemi di NLP è una tendenza che senza dubbio rimarrà parte integrante dei metodi contemporanei nel prossimo futuro.

1.3 APPLICAZIONI DI ML E NLP ALLA SALUTE MENTALE

Le applicazioni del ML e della PNL alla salute mentale possono essere classificate seguendo queste due direttrici:

- Il corpus: poiché uno degli argomenti è la PNL, il corpus ha necessariamente una componente testuale. I corpora più comuni sono documenti o rapporti (cartelle cliniche elettroniche [EHR], rapporti di valutazione psicologica e rapporti del medico legale), social media (Reddit, Twitter, ecc.) o interviste trascritte a pazienti.

- Elaborazione del corpus: a seconda della natura del corpus, si possono estrarre i termini medici e abbinarli agli identificatori univoci di concetto (CUI) del sistema unificato di linguaggio medico (UMLS) oppure elaborare blocchi di testo in linguaggio naturale ed eseguire ricerche specifiche (ad esempio, per individuare i termini relativi al suicidio).
- Metodi di classificazione: vengono utilizzate molte tecniche di ML, come alberi decisionali, macchine a vettori di supporto, campi casuali condizionati, Random Forest e NN.
- Obiettivo: l'obiettivo è solitamente quello di convalidare un'ipotesi o di studiare il comportamento di una determinata popolazione di pazienti.

I corpora possono essere molto voluminosi. Ad esempio, Sinnenberg et al [36] hanno pubblicato una revisione sistematica su Twitter come strumento per la ricerca sanitaria, che ha incluso 137 studi diversi e analizzato oltre 5 miliardi di tweet utilizzando il metodo ML; Castro et al [37] hanno elaborato 4,2 milioni di EHR per un periodo di oltre 20 anni. I corpora possono anche essere piccoli, come dimostra lo studio condotto da Carson et al [38], che ha trattato le note cliniche non strutturate di 73 intervistati, o lo studio di Bedi et al [39], in cui sono state considerate solo le interviste narrative di un'ora di 34 partecipanti. A volte, i corpora vengono creati appositamente per un progetto. Ad esempio, in uno studio di Roy et al [40], dei volontari hanno scritto 182 testi con caratteristiche di abuso, annotati da ricercatori e vittime di abusi; questi testi sono stati poi analizzati e hanno fornito un modello per individuare i testi con tratti di abuso.

L'estrazione delle CUI dell'UMLS si applica principalmente ai documenti EHR perché questi ultimi sono semistrutturati e costituiscono un tipo di documento speciale. Le specificità di questo tipo di documento si riflettono nella sua struttura, nella sintassi del testo e, soprattutto, nel vocabolario utilizzato. L'estrazione dei termini medici avviene attraverso algoritmi di estrazione dell'informazione e la corrispondenza di questi termini con i CUI dell'UMLS viene effettuata attraverso metodi di rappre-

sentazione della conoscenza. Una volta estratti i concetti da un EHR, quest'ultimo viene rappresentato dai primi e i concetti diventano caratteristiche utilizzate per la classificazione.

Su corpora diversi dagli EHR, invece di estrarre i CUI dell'UMLS, si applicano metodi NLP più generali ai dati testuali per ottenere caratteristiche che vengono poi classificate da algoritmi di ML. Questi metodi NLP sono spesso costituiti da conteggi di frequenza di parole o *n-grammi* in un insieme specifico, che può essere curato manualmente o ottenuto da un corpus. In altri casi, per l'individuazione dei temi vengono utilizzati metodi come LSA o l'allocazione di Dirichlet latente (LDA). L'insieme iniziale di parole può essere esplicito. Ad esempio, Doan et al [41] hanno raccolto i tweet contenenti gli hashtag #relax e #stressed e li hanno classificati per tema e luogo. In altri casi, i calcoli vengono eseguiti a un livello superiore e le parole coinvolte nel processo non sono esplicitamente note. Ad esempio, Luo et al. [42] hanno cercato di caratterizzare l'autismo analizzando le descrizioni testuali di individui strettamente correlati scritte da pazienti o membri di un gruppo di controllo. Tuttavia, la maggior parte delle applicazioni NLP per la salute mentale si basa sulle parole (utilizzando il metodo del bag-of-words [borsa di parole], cioè ignorando l'ordine delle parole e mantenendo solo la frequenza con cui vengono utilizzate). Alcune tengono conto dell'ordine delle parole in modo limitato (utilizzando gli *n-grammi*, cioè sequenze contigue di parole di lunghezza *n*), ma pochissime tengono conto della sintassi utilizzando gli alberi di dipendenza [18,43,44]. Per quanto riguarda le loro applicazioni, va notato che gli strumenti di ML e NLP sono preziosi per limitare i problemi di dati come l'eccesso di dati presente nella medicina moderna. Forsting et al. [45] riconoscono che le tecniche di ML e NLP possono essere utili per l'optimism bias (ad esempio, la differenza tra le aspettative di una persona e i risultati effettivi oppure il pensiero di un medico che crede che il problema del suo paziente ricada nel proprio settore di specializzazione) perché la macchina ha un approccio generalista a differenza del medico specialista. Negli ultimi due decenni, queste tecniche sono

emerse nel campo della salute mentale, in seguito al successo dei social media come fonte informativa di dati [46].

Inoltre, la PNL è essenziale in psichiatria perché i deficit linguistici sono sintomi comuni di depressione, disturbi comportamentali, DSA, disturbi di personalità e schizofrenia [30]. Può fornire informazioni sulla salute mentale e sulle emozioni degli individui, sull'uso di stili di linguaggio narrativi, soggettivi e strutturati e sul loro stile di vita, in particolare sul livello di istruzione, sullo stato socioeconomico, sulle condizioni di vita e sull'ambiente culturale [47], tutti elementi che sono di routine negli esami dello stato mentale.

Utilizzando il ML in generale e i metodi NLP in particolare, si possono creare sistemi semiautomatici (che operano sotto la supervisione umana) con l'obiettivo di migliorare la specificità della diagnosi, la conoscenza della psicofisiologia, la velocità della diagnosi e stime più accurate della gravità della malattia [48]. Attraverso l'analisi dei post su Twitter, O'Dea et al. [49] hanno individuato l'importanza di creare campagne in tempo reale per spingere a cercare aiuto e ridurre lo stigma legato alla salute mentale. Inoltre, i programmi automatizzati possono essere più efficienti in termini di costi e di tempo rispetto alle loro controparti tradizionali. Ly et al. [50] hanno proposto di utilizzare interventi basati su un sistema di auto-aiuto automatizzato come modo per rendere più accessibili gli strumenti di promozione della salute mentale. Inoltre, Lucas et al. [51] hanno dimostrato, attraverso uno studio clinico, che quando le persone credevano di interagire con un computer piuttosto che con un medico vero e proprio riferivano di avere meno paura di rivelare se stesse, riducevano i comportamenti di autocontrollo, provavano maggiore facilità nell'esprimere la gravità delle loro emozioni e venivano valutate dagli osservatori come più disposte ad aprirsi. Tuttavia, questi risultati potrebbero non essere generalizzabili, in quanto potenzialmente influenzati dalla selezione del campione e/o dalla progettazione del sistema stesso.

Sebbene il ML e l'NLP forniscano nuovi strumenti e strategie per la ricerca e la pratica psichiatrica [52], occorre tenere

presente che il loro uso solleva spesso problemi etici e legali sul consenso all'uso dei dati personali e sull'anonimizzazione dei dati. Allo stesso modo, gli studi che utilizzano l'IA per le analisi predittive mettono in discussione l'equilibrio tra benefici apportati e rispetto dell'autonomia dei pazienti. McKernan et al. [53] suggeriscono di impegnarsi a comunicare i metodi di IA per ottenere il consenso libero e informato dei pazienti. Inoltre, dovrebbero essere condotti studi prospettici per valutare l'uso degli strumenti di IA [53].

L'obiettivo primario di questa revisione sistematica è riassumere e catalogare gli studi che hanno utilizzato tecniche di ML e NLP per la salute mentale in termini metodologici e tecnici. L'obiettivo secondario è quindi quello di considerare l'uso potenziale di questi metodi nella pratica clinica della salute mentale, come i contributi che possono offrire nelle aree della diagnosi e della prognosi, la definizione dei fattori di rischio, l'impatto della psicoterapia, l'adeguatezza al trattamento e gli effetti collaterali.

2. METODI

Questa revisione sistematica si basa sulle linee guida PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analysis) [54]. Le ricerche sono state condotte come specificato dal protocollo standard di PROSPERO (Prospective Register of Systematic Reviews; numero di registrazione CRD42019107376).

2.1 STRATEGIA DI RICERCA DELLA LETTERATURA

È stata condotta una ricerca sistematica e computerizzata della letteratura utilizzando 4 banche dati: PubMed (via MEDLINE), Scopus, ScienceDirect e PsycINFO. Ogni database è stato esplorato dal 21 agosto 2018 al 1° febbraio 2020, senza limiti di data di pubblicazione. La ricerca è stata effettuata utilizzando le seguenti parole chiave: "natural language processing" AND "machine learning" AND ("psychiatry" OR "mental health" OR "mental disorder"). La stessa ricerca è stata eseguita sull'elemen-

to (data mining) invece che su (machine learning). Quando il testo completo non era disponibile, è stato utilizzato l'abstract per estrarre le informazioni necessarie per evitare bias di selezione. Sono stati esclusi i casi di studio, i documenti di conferenze e le revisioni.

2.2 SELEZIONE DELLO STUDIO E CRITERI DI AMMISSIBILITÀ

Dopo aver eliminato i duplicati, 2 collaboratori hanno esaminato in modo indipendente tutti i titoli e gli abstract rilevanti per questa revisione sistematica. Un terzo revisore è stato consultato in caso di disaccordo tra i primi due collaboratori. Il processo è illustrato nell'Appendice multimediale 1. Sono stati selezionati solo gli studi disponibili in lingua inglese. Abbiamo deliberatamente escluso gli studi sul processo di anonimizzazione per concentrarci sugli articoli che indagano l'uso clinico della ML e della PNL in psichiatria (ad esempio, il contributo alla diagnosi, alla prognosi, alla definizione dei fattori di rischio, all'impatto della psicoterapia, all'aderenza al trattamento e agli effetti collaterali). Non sono state applicate limitazioni sulle date di pubblicazione. In totale sono stati inclusi nella revisione 58 articoli.

2.3 STUDI INCLUSI

Tutti gli studi sono stati accuratamente vagliati e le loro idee principali sono riassunte in singole tabelle (Appendice multimediale 2 [37-41,43,47,48,55-104]). Queste tabelle forniscono informazioni sulle caratteristiche qualitative e quantitative: autori, anno di pubblicazione, argomento preciso della salute mentale (ad esempio, autismo, disturbo dello spettro psicotico, ecc.), caratteristiche della popolazione, tipi e volume dei dati registrati. La seconda parte di queste tabelle riassume gli obiettivi, i metodi e i risultati.

3. RISULTATI

3.1 SELEZIONE DELLO STUDIO

La ricerca nel database ha permesso di identificare 222 studi utilizzando la parola chiave (machine learning) e 105 studi utilizzando la parola chiave (data mining). Dopo averli uniti, sono stati considerati 238 studi unici per la revisione, sulla base del titolo e dell'abstract. Un totale di 84 articoli sono stati esclusi perché (1) non riguardavano la psichiatria o la salute mentale (52 casi), (2) non erano scritti in inglese (1 caso) e (3) le parole chiave (machine learning), (natural language processing) o (data mining) non comparivano nel titolo o nell'abstract (8 casi). Come secondo filtro, sono stati esclusi 33 studi sull'anonimizzazione dei dati. Inoltre, 7 studi sono stati esclusi perché il ML o il NLP non erano l'argomento principale, ma erano solo citati come informazioni collaterali. Inoltre, 96 articoli sono stati esclusi perché si trattava di revisioni, casi di studio o documenti di conferenze. Infine, 58 articoli sono stati inclusi in questa revisione.

3.2 TEMI E POPOLAZIONE

Gli argomenti sono eterogenei. Gli argomenti più frequentemente citati sono la depressione e il suicidio con 17 studi [38,55,57,60-62,77-79,82,83,87,88,91,92,99,104]. Altre diagnosi psichiatriche sono state la dipendenza da alcol o droghe illecite (6 casi) [43,65,66,75,84,86]; il disturbo post-traumatico da stress (PTSD; 3 casi) [47,63,64]; i disturbi del neurosviluppo (3 casi) [42,58,93]; disturbi dello spettro psicotico, inclusa la schizofrenia (3 casi) [39,95,100]; ansia (2 casi) [41,98]; disturbo di personalità (1 caso) [85]; disturbi alimentari (2 casi) [89,96]; e disturbo bipolare (2 casi) [37,102]. In totale, 3 studi riguardavano la violenza e le molestie informatiche [40,80,94]. Sono stati inoltre descritti problemi di trattamento come l'aderenza o l'abuso (6 casi) [56,72,74,81,101,103]. Sono stati trovati solo 1 studio sulla contenzione fisica [90] e 1 sui disturbi cognitivi [97]. In totale,

8 studi erano transnosografici [59,67-71,73,76]: 6 soddisfacevano i criteri del CEGS N-GRID 2016 Center of Excellence in Genomic Science Neuropsychiatric-Genome-Scale and Research Domain Criteria (RDoC) Individualized Domains 2016 Shared Task in Clinical NLP, che saranno sviluppati ulteriormente nei nostri risultati.

In totale, sono state individuate 3 categorie distinte di popolazione:

1. Pazienti le cui cartelle cliniche elettroniche erano disponibili in database di ricerca basati su dati scientifici, come la cartella clinica elettronica (EMR) di Partners HealthCare, una raccolta di dati di pazienti del Massachusetts General Hospital e del Brigham and Women's Hospital [55,56]. Questi dati andavano oltre le cartelle psichiatriche e includevano anche altre cartelle cliniche.
2. Pazienti visitati nei reparti di emergenza o di psichiatria che presentavano caratteristiche cliniche aggiuntive nella loro cartella (ad esempio, osservazione clinica, esami di laboratorio, interventi diagnostici e terapeutici, note dattiloscritte degli specialisti).
3. Reti di social media (Facebook, Twitter e Instagram): Gli autori di questi studi hanno selezionato hashtag specifici come #stress o #depressione e hanno vagliato una moltitudine di messaggi pubblici utilizzando una piattaforma di streaming.

3.3 OBIETTIVI

In totale, sono state individuate 5 categorie principali di obiettivi: estrarre i sintomi clinici, classificare la gravità delle malattie, confrontare le diverse terapie, fornire indizi psicopatologici nella salute mentale e sfidare la nosografia attuale.

Gli obiettivi principali di questi studi erano estrarre e registrare i sintomi clinici, stabilire una diagnosi o monitorare i cam-

biamenti nel tempo. In totale, 2 studi si occupavano di monitoraggio epidemiologico automatizzato: Metzger et al [57] hanno fornito un metodo per rilevare i tentativi di suicidio dalle EHR e Leroy et al [58] hanno ottenuto l'estrazione automatica dei criteri per l'ASD dalle EHR con un'accuratezza del 76%. Quest'ultimo studio ha affermato che dal 2000 al 2010 si è verificato un aumento della prevalenza di determinati sintomi (comportamento non verbale, reciprocità sociale ed emotiva e aderenza alle disabilità di routine). L'estrazione dei dati è stata utilizzata anche per la diagnosi: He et al [47] hanno diagnosticato il PTSD con un'accuratezza dell'82% dopo aver analizzato testi liberi scritti da sopravvissuti al trauma.

Oltre all'estrazione, un obiettivo importante era quello di misurare la gravità dei disturbi psichiatrici nei corpora di valutazione psicologica. Goodwin et al [59] hanno classificato i sintomi dei pazienti con psicosi in 4 diversi livelli di gravità (assente, lieve, moderata e grave) utilizzando analisi statistiche. Fernandes et al. [60] hanno studiato le cartelle cliniche elettroniche di una coorte di individui con una storia di tentativi di suicidio e di una coorte di individui con una storia di sola premeditazione di suicidio. Il loro algoritmo per rilevare la premeditazione suicida o i tentativi di suicidio aveva una sensibilità del 98,2% e un valore predittivo positivo dell'82,8% [57]. Altri studi hanno riscontrato che le tecniche di ML e NLP hanno dato buoni risultati, anche se non erano necessariamente migliori della capacità di un medico di prevedere il rischio clinico di suicidio nei propri pazienti [61,62]; pertanto, gli autori hanno proposto approcci statistici NLP da utilizzare in parallelo con la pratica clinica.

I metodi ML e NLP sono utilizzati anche per misurare e confrontare l'efficacia di diversi tipi di psicoterapia [63,64]. Tanana et al. [43] hanno studiato 2 tecniche statistiche di PNL per codificare le sessioni di colloquio motivazionale. Il colloquio motivazionale è un metodo di psicoterapia utilizzato per i disturbi da uso di sostanze e altri problemi comportamentali, per rafforzare la motivazione personale al cambiamento [105]. I colloqui motivazionali possono essere codificati manualmente per valutare l'a-

derenza alla terapia e raccogliere feedback per le sessioni successive. Gli autori hanno riscontrato che il modello di caratteristiche discrete delle frasi (un classificatore di frasi basato su modelli di *n-grammi*) aveva un'accuratezza simile alla procedura tradizionale delle sessioni terapeutiche. Maguen et al. [63] hanno utilizzato tecniche statistiche di NLP per distinguere la psicoterapia basata sull'evidenza, come la terapia di elaborazione cognitiva e le note di esposizione prolungata, dalle note di psicoterapia non strutturate per una popolazione di veterani con PTSD. Hanno scoperto che quasi il 20% dei veterani ha osservato un miglioramento dei sintomi dopo una o più sedute di psicoterapia basata sull'evidenza.

Un altro obiettivo era quello di fornire indizi psicopatologici per la comprensione dei disturbi mentali attraverso l'analisi delle caratteristiche del linguaggio. Questo obiettivo comporta talvolta l'elaborazione di dati precedentemente inesplorati, come i discorsi in gruppo o le reti sociali. Di seguito sono riportati alcuni esempi di studi che perseguono questo obiettivo: Baggott et al [65] hanno scoperto che l'MDMA (3,4-méthylènedioxy-N-méthylamphétamine; Ecstasy) alterava i modelli di discorso degli individui più frequentemente rispetto al placebo e portava a un aumento dell'uso del linguaggio sociale e sessuale sia positivo che negativo (altri, pubblico, cameratismo ed estroverso). Chary et al. [66] hanno analizzato i post su Lycaeum, un popolare forum web, una delle piattaforme più citate quando si parla di uso di droghe. Qui hanno scoperto nuove combinazioni di farmaci non citate nella letteratura medica. Luo et al. [42] hanno differenziato le interazioni sociali tra adulti con DSA e adulti sani. Hanno confermato l'ipotesi relativa alle differenze nel linguaggio e nelle interazioni sociali negli adulti con DSA: i partecipanti tipici avevano più collegamenti semantici rispetto al gruppo DSA e le parole con il maggior numero di collegamenti erano diverse tra i due gruppi. Doan et al. [41] hanno notato che gli utenti americani di Twitter sono più propensi a esprimere la loro fonte di stress su Twitter che nelle loro esperienze quotidiane. Le principali cause di stress emerse dai dati di Twitter sono state

l'istruzione, il lavoro e le relazioni sociali. Hanno anche scoperto che le espressioni di stress e comfort degli individui differiscono in base alla città di residenza (Los Angeles, New York, San Diego e San Francisco). Inoltre, Mowery et al. [106] hanno rivelato che meno del 2% del corpus di tweet (un corpus di 9300 tweet annotati contenenti parole chiave relative alla depressione) includeva più di un riferimento alla depressione, suggerendo che possono esistere diversi canali espressivi quando si parla di depressione.

Infine, l'IA nella ricerca sulla salute mentale sfida la pratica e la nosografia attuali. Nel 2010, Insel et al [107] hanno avviato un progetto chiamato RDoC, un quadro di ricerca per i disturbi mentali che mira a costituire un'alternativa al DSM (Manuale diagnostico e statistico dei disturbi mentali). Il primo include dati di genetica e neuroscienze nella sua classificazione dei disturbi mentali, il secondo invece si basa esclusivamente su dati clinici [107]. La RDoC è una matrice in cui le colonne e le righe rappresentano i costrutti (geni, molecole, cellule, circuiti, fisiologia, comportamenti, auto-rapporti e paradigmi) e i sotto-costrutti di ciascuno dei seguenti 6 domini: valenza negativa, valenza positiva, sistemi cognitivi, sistemi per i processi sociali, sistemi di eccitazione o di regolazione e sistemi sensomotori. I fautori della RdC sostengono che le sindromi del DSM presentano limitazioni significative quando vengono utilizzate come fenotipi per identificare biomarcatori e varianti genetiche specifiche associate alla malattia mentale [108]. Un'applicazione concreta di questo nuovo sistema ha utilizzato metodi statistici NLP per creare una coorte fenotipicamente omogenea che ha permesso un migliore confronto [109]. Nel 2016, il CEGS N-GRID (Centers of Excellence in Genomic Science Neuropsychiatric-Genome-Scale and RDoC Individualized Domains) ha proposto 3 compiti impegnativi utilizzando metodi NLP: (1) l'anonimizzazione dei dati, (2) la previsione della gravità dei sintomi nel dominio della valenza positiva dalle cartelle cliniche neuropsichiatriche e (3) un nuovo caso d'uso dei dati (ad esempio, la previsione della presenza di condizioni mentali comuni nei pazienti) [67]. Questa ricerca sull'elaborazione NLP e ML ha identificato 6 articoli [59,67-71]

che soddisfano questi compiti di sfida, anche se solo uno studio si è occupato del compito 3 [67]. Come accennato in precedenza, sono stati esclusi gli studi sull'anonimizzazione; pertanto, il framework RDoC collega le basi neuro-biologiche dei processi mentali con le manifestazioni fenotipiche [110]. Il compito condiviso CEGS N-GRID ha fornito dati utilizzabili per lo studio di tecniche ML e NLP, che potrebbero portare a nuove nosologie psichiatriche.

3.4 TIPO DI DATI UTILIZZATI

Come si può vedere nella Tabella 1 (in cui non sono visualizzati gli unicum), il tipo di corpus più frequente è quello degli EHR (a cui si possono aggiungere gli EMR). Gli EHR (e gli EMR) sono fonti di dati ottimi quanto a eterogeneità: combinano dati strutturati, semistrutturati e liberi e spesso utilizzano un linguaggio molto controllato contenente termini medici che consentono l'estrazione di CUI (sezione Metodi). Le seconde fonti di dati più frequenti sono le note cliniche e le cartelle cliniche, che condividono le proprietà degli EHR o degli EMR, ma non sono standardizzate allo stesso modo.

Caratteristiche	Valori
EHRsa	22.9508
Note cliniche	16.3934
Registri clinici	11.4754
Interviste	8.1967
Tweet	8.1967
Questionari	6.5574
Reddit	6.5574
Web	4.918
EMRsb	3.2787

Tabella 1. Tipo di corpus. a) EHR: cartella di salute elettronica. b) EMR: cartella medica elettronica.

I dati descritti in precedenza condividono un'importante proprietà: i corpora sono generati da professionisti e quindi possono essere utilizzati per l'estrazione di termini medici con risultati soddisfacenti.

Un'altra categoria di dati è quella generata dai pazienti. Questa categoria può essere suddivisa in due sottocategorie: i dati generati con l'aiuto degli operatori (ad esempio, interviste e questionari) e i dati generati liberamente dai pazienti sui social media (tweet, post su Reddit, blog web, ecc.).

Le interviste e le parti testuali dei questionari sono tecnicamente testi liberi, ma gli operatori hanno comunque un certo controllo sul contenuto e l'ambiente in cui vengono raccolti i dati influenza il grado di informalità dei testi. Per questi motivi, i metodi tradizionali di PNL possono essere applicati ad essi con risultati soddisfacenti.

I dati raccolti dai social media, a causa dell'alto grado di informalità, dell'ortografia e della sintassi poco rigorose e dell'uso di abbreviazioni ed emoji, possono essere elaborati solo superficialmente dai metodi NLP standard. Esempi tipici sono gli studi di Doan et al [41] e Jackson et al [73], in cui i tweet sono stati selezionati perché contenevano gli hashtag #stress e #relax e le loro parole sono state utilizzate in un bag-of-words senza alcun ulteriore trattamento linguistico [41] o i tweet sono stati selezionati in base alla presenza di termini che indicano gli oppioidi [72]. Sebbene gli autori abbiano lemmatizzato i contenuti dei tweet, la caratteristica principale dei tweet presi in considerazione è stata la loro origine geografica.

3.5 METODOLOGIA

Sono state distinte due fasi dei progetti NLP: (1) la preelaborazione, che consiste nell'analisi dei dati per ottenere caratteristiche numeriche o categoriali, e (2) la classificazione.

3.6 PREELABORAZIONE

La tabella 2 (in cui non sono visualizzati gli apici) rappresenta la frequenza d'uso di vari metodi di pre-elaborazione, che possono essere di natura diversa. Alcuni metodi si applicano a parole o gruppi di parole: *lemma* (lemmatizzazione, cioè sostituzione di una parola con una forma base come il singolare per i nomi o l'infinito per i verbi), *POS* (part of speech, cioè attribuzione a una parola di un'etichetta che ne denota la funzione grammaticale), *cTAKES* o *CUI* (mappatura di una parola o di una frase che sottosta a un'ontologia, come l'UMLS, e quindi definizione univoca della sua semantica), *tf-idf* (attribuire a una parola o a un termine un valore che ne rappresenti il significato nella caratterizzazione di un dato documento o di una classe a cui appartiene), *embedding* (rappresentare una parola con un vettore in uno spazio ad alta dimensionalità), *named-entity recognition* (decidere se una data parola o frase sostantiva è un'entità denominata), *LIWC* (Linguistic Inquiry and Word Count, uno strumento commerciale pubblicizzato come "basato su solide basi scientifiche" che fornisce varie "intuizioni sociali e psicologiche" delle parole). Altri metodi combinano le parole in strutture più elevate: gli *n-grammi* (l'*n-gramma*, cioè una sequenza di *n* parole successive, è considerato come un'unità di cui viene misurata la frequenza). Infine, altri metodi vengono applicati a intere frasi, paragrafi o documenti: *SentiAna* (che analizza i sentimenti o le emozioni), *LDA* e *LSA* (che calcolano insiemi di argomenti, individuano la significatività di ogni argomento per un dato documento e forniscono parole rappresentative per ogni argomento). I metodi di preelaborazione più frequenti sono i metodi standard di NLP (lemmatizzazione, part-of-speech tagging, *n-grammi* e tf-idf) e metodi specifici per i testi medici, come l'estrazione di CUI (parole chiave cTAKES e CUI nella Tabella 2). Il metodo dell'embedding è legato quasi esclusivamente alle NN e quindi è relativamente recente. Infine, la coda del grafico nella Tabella 2 contiene metodi applicati principalmente ai testi liberi, come il

rilevamento di argomenti, il riconoscimento di entità denominate, la sentiment analysis o l'analisi delle emozioni.

Caratteristiche	Valori
lemma	16.3043
POSa	10.8696
cTAKESb	10.8696
ngrammi	9.7826
tfidf	7.6087
incorporazione	6.5217
CUIsc	5.4348
LDAd	5.4348
SentiAna	5.4348
LIWCe	4.3478
NERf	4.3478
LSAg	3.2609

Tabella 2. Metodi di pre-elaborazione.
- *a) POS: parte del discorso [Part Of Speech].*
- *b) cTAKES: sistema clinico di analisi del testo e di estrazione della conoscenza [clinical Text Analysis and Knowledge Extraction System].*
- *c) CUI: identificatore univoco del concetto [Concept Unique Identifier].*
- *d) LDA: allocazione di Dirichlet latente [Latent Dirichlet Allocation].*
- *e) LIWC: Inchiesta linguistica e conteggio delle parole [Linguistic Inquiry and Word Count].*
- *f) NER: riconoscimento di entità nominali [Name Entity Recognition]]*
- *g) LSA: analisi semantica latente [Latent Semantic Analysis].*

3.7 CLASSIFICAZIONE

Una volta raggiunta la fase di classificazione, i dati linguistici vengono interamente convertiti in dati numerici e, pertanto, la scelta del classificatore dipende da fattori che variano a seconda del tipo di corpus. Alcuni di questi fattori includono (1) il volume dei dati, (2) il tipo di classificazione (supervisionata o non

supervisionata), (3) il livello di spiegabilità e (4) la piattaforma utilizzata. Nella Tabella 3 (dove gli apici non sono visualizzati), abbiamo mostrato l'albero decisionale, le regole di associazione e C4.5 (anch'esso un algoritmo ad albero decisionale) che sono metodi *trasparenti*, cioè l'utente può seguire il processo di classificazione in modo graduale e capire il motivo per cui un dato individuo appartiene a una particolare classe. Non sono i classificatori più frequenti, probabilmente perché la spiegabilità non è una delle principali preoccupazioni della maggior parte degli studi. Invece, i classificatori più frequentemente utilizzati, come la Support Vector Machine (SVM), la LogiR (regressione logistica), la RF (Random Forest) e la LinR (regressione lineare), sono classificatori solidi e veloci, con piccoli set di parametri e buone prestazioni. Al centro della Tabella 3 si trovano i NN che appartengono alla tendenza dell'apprendimento profondo della ML: sono opposti a DT/AR/C4.5 quando si tratta di spiegabilità e si basano molto su alcuni parametri (tipo e geometria del NN, numero di strati, dimensione degli strati, ottimizzatore, tasso di apprendimento, funzione di perdita, ecc.) Le cause della frequenza relativamente bassa delle NN nelle pubblicazioni possono essere (1) il fatto che sono state implementate in framework di facile utilizzo (come Theano o Keras) solo di recente, (2) la necessità di mettere a punto un gran numero di parametri e (3) i requisiti relativamente elevati in termini di memoria, di elaborazione centrale e grafica. È probabile che questa situazione cambi nel prossimo futuro.

Caratteristiche	Valori
SVMa	22.6804
LogiRb	16.4948
RFc	11.3402
DTd	6.1856
NBe	6.1856

NNf	6.1856
LinRg	5.1546
K-Means	3.0928
ARh	2.0619
C4.5	2.0619

Tabella 3. Tipo di classificatore.
- *a) SVM: support vector machine.*
- *b) LogiR: regressione logistica [logistic regression].*
- *c) RF: random forest.*
- *d) DT: albero decisionale [decision tree].*
- *e) NB: Naive Bayes.*
- *f) NN: rete neurale [neural network].*
- *g) LinR: regressione lineare [linear regression].*
- *h) AR: regole di associazione [association rules].*

3.8 PIATTAFORME

Come si può vedere nella Tabella 4 (gli apici non sono rappresentati), le due piattaforme più comuni sono Python e R. Python è un linguaggio di programmazione *universale*, nel senso che non è specifico per un determinato dominio: i suoi oltre 120.000 pacchetti permettono all'utente di eseguire compiti specializzati in qualsiasi campo possibile. Inoltre, è open-source e la documentazione di alta qualità è abbondante. Anche R è un linguaggio di programmazione e un compilatore open-source, ma, contrariamente a Python, è orientato alla statistica. Sebbene molti classificatori siano stati implementati in modo efficiente sia in Python che in R, il dominio dell'NLP è meglio rappresentato in Python, grazie a pacchetti come NLTK (Natural Language ToolKit), spaCy e Stanza. La terza riga ("Sconosciuto") rappresenta le pubblicazioni che non menzionano la piattaforma utilizzata. La quarta riga indica la piattaforma General Architecture for Text Engineering General Health, un'applicazione Java open-source che fornisce un ambiente per l'elaborazione di dati testuali in modo semplice. La riga *Apache* raccoglie diversi strumenti distribuiti dalla Apache Software Foundation. Stata è un software statistico

commerciale di College Station, Texas, rilasciato per la prima volta nel 1985. Weka è un ambiente di programmazione open-source per il ML.

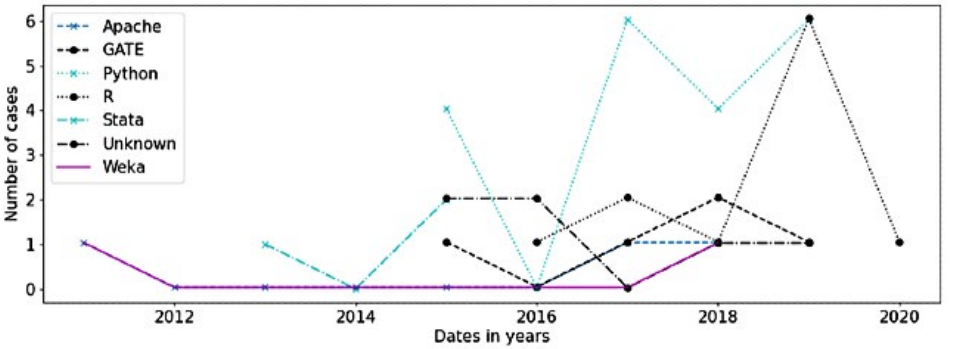


Figura 1. Utilizzo delle piattaforme. Visualizza questa figura

La Figura 1 mostra l'uso delle piattaforme in ordine cronologico. L'uso di Python e R è iniziato dopo il 2015, mentre Stata, Weka e Apache erano già in uso nel 2011.

Caratteristiche	Valori
Python	34.4828
R	18.9655
Sconosciuto	10.3448
GATEa	8.6207
Apache	5.1724
Stata	5.1724
Weka	3.4483

Tabella 4. Piattaforme. aGATE: Architettura generale per l'ingegneria del testo Salute generale.

3.9 ANALISI DELLE CORRISPONDENZE TRA DATI, METODI, CLASSIFICATORI, PIATTAFORME E PUBBLICAZIONI

L'analisi delle corrispondenze è una tecnica di riduzione delle dimensioni che mappa i dati in uno spazio fattoriale dove ogni dimensione è una combinazione delle variabili iniziali.

La figura 2 rappresenta le coordinate principali delle pubblicazioni e le varie entità considerate nel loro studio.

A destra, un gruppo di pubblicazioni è circondato dal tipo di dati *ClinNotes*, dal metodo *cTAKES* e dalla piattaforma *R*. Nel quadrante superiore sinistro, alcune pubblicazioni si riuniscono con il metodo *embedding* e il classificatore *NN*. Verso la sinistra del diagramma e vicino all'asse orizzontale, sono presenti pubblicazioni con una piattaforma *sconosciuta* che utilizza il classificatore *NB* e un grande cluster il cui centro comprende *tf-idf*, *LogiR*, *SVM*, *Python* e *n-grammi*: i sistemi legacy, i classificatori più utilizzati e la piattaforma più utilizzata.

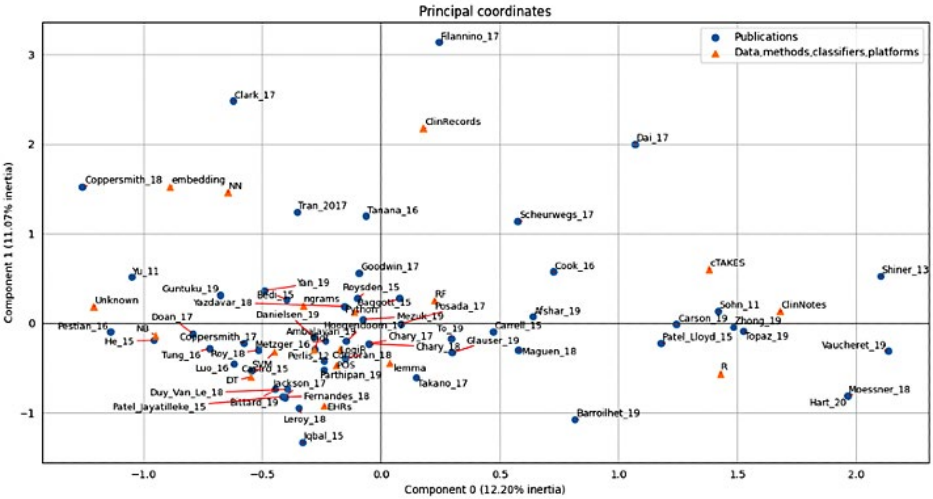


Figura 2. Analisi delle corrispondenze. Visualizza questa figura

Per quanto riguarda le pubblicazioni, *Filannino_17* è un ovvio outlier perché non ha un metodo, un classificatore o una piat-

taforma e perché descrive un compito e come questo compito è stato trattato da altri. *Clark_17* si trova all'estremo superiore sinistro, poiché utilizza NN e k-means (quest'ultimo non viene visualizzato perché vengono incluse solo le entità che compaiono almeno 5 volte). Anche *Coppersmith_18* utilizza embeddings e NN, mentre *Tran_17* (più vicino al cluster centrale) utilizza sia NN che SVM. A destra, *Shiner_13* e *Vaucheret_19* utilizzano note cliniche e R, mentre *Hart_20* e *Moessner_18* utilizzano R e metodi che non sono stati presi in considerazione nel calcolo. In basso a sinistra, *Iqbal_15* utilizza EHR nell'architettura generale per l'ingegneria testuale della salute (che non viene visualizzata). All'estrema sinistra e vicino all'asse orizzontale, *Pestian_16* e *Yu_11* utilizzano una piattaforma sconosciuta.

3.10 DISTRIBUZIONE GEOGRAFICA DEGLI AUTORI

Nella mappa in Figura 3, il diametro dei segni rossi è proporzionale a un punteggio calcolato come segue: abbiamo aggiunto 1 unità per l'origine geografica dell'affiliazione di ciascun autore di ogni articolo. Le città con punteggi superiori a 10 sono Boston (54), Londra (44), New York (21), Cincinnati (15), Buenos Aires (13), Cambridge, Massachusetts (12), San Francisco (11) e Taiwan (11).

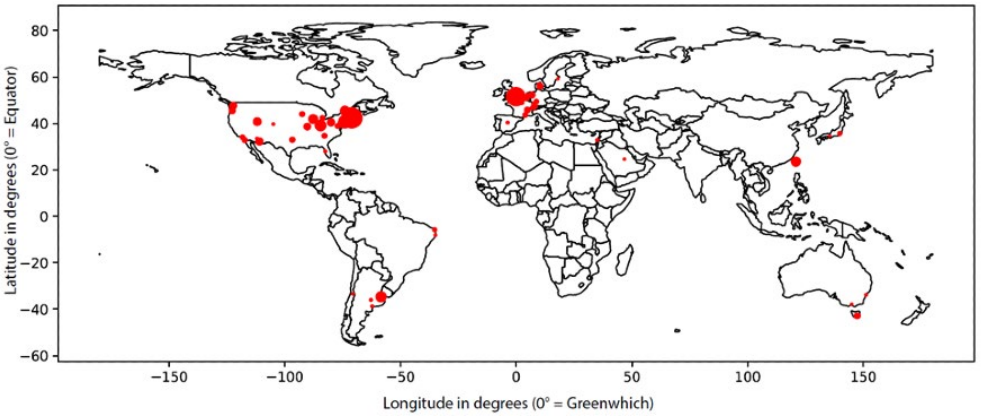


Figura 3. Distribuzione geografica degli autori. Visualizza questa figura

3.11 CITAZIONI E COCITAZIONI

La Figura 4 rappresenta le citazioni degli articoli della nostra lista da parte di altri articoli della stessa lista. La dimensione dei nodi di un articolo è proporzionale al numero di articoli che lo citano. I colori dei nodi e dei bordi rappresentano le comunità. Ogni comunità ha un nodo centrale: Perlis et al [55] sono citati in altri 7 articoli, Jackson et al [73] sono citati in altri 4 articoli, Carrell et al [74] e Afshar et al [75] sono citati in altri 2 articoli e Bedi et al [39] sono citati in altri 2 articoli. In totale, 22 articoli sono singolari: non sono citati né citano altri articoli nel nostro elenco.

Sebbene le citazioni reciproche mostrino le influenze tra gli articoli della nostra lista, possiamo anche misurare il numero di cocitazioni (cioè i riferimenti comuni tra due articoli della lista). Nella Figura 5, i bordi tra gli articoli indicano che hanno almeno 3 riferimenti comuni. L'ampiezza dei bordi è proporzionale al numero di riferimenti.

Il bordo di maggiore ampiezza è quello tra i lavori di Coppersmith et al [76] e Coppersmith et al [77], il che è normale: i due lavori hanno lo stesso primo autore, sono stati pubblicati in meno di un anno e hanno 26 riferimenti comuni.

Il secondo caso, in ordine decrescente di ampiezza dei bordi, è tra Shiner et al [64] e Maguen et al [63]. Anche questo è normale: il primo autore del primo è anche l'ultimo autore del secondo e quest'ultimo è presentato come un'estensione del primo: "In questo studio, il nostro obiettivo è stato quello di estendere il lavoro di Shiner e colleghi applicando la codifica automatica a un ampio pool nazionale di procedure di trattamento della salute mentale al fine di identificare l'uso della terapia di elaborazione cognitiva e dell'esposizione prolungata". I due lavori hanno in comune 14 riferimenti.

La dimensione dei nodi nel grafo è proporzionale al grado. Zhong et al [78] hanno il grado più alto: questo articolo ha più di tre riferimenti comuni con ben otto altri articoli, anzi, con 8 riferimenti. Il colore dei nodi e degli spigoli corrisponde alle

componenti connesse. Ci sono 19 nodi singoli che condividono ≤2 riferimenti con ogni altro articolo della lista.

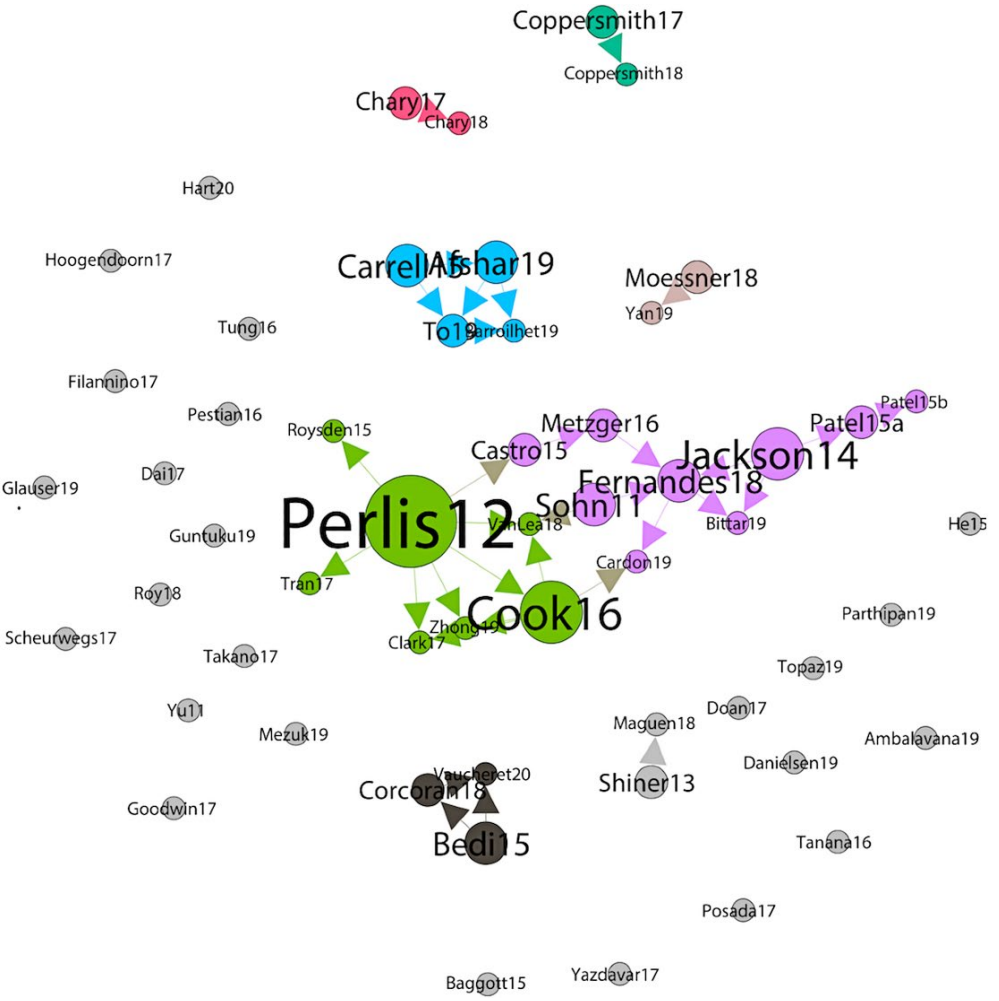


Figura 4. Grafico delle citazioni. Visualizza questa figura

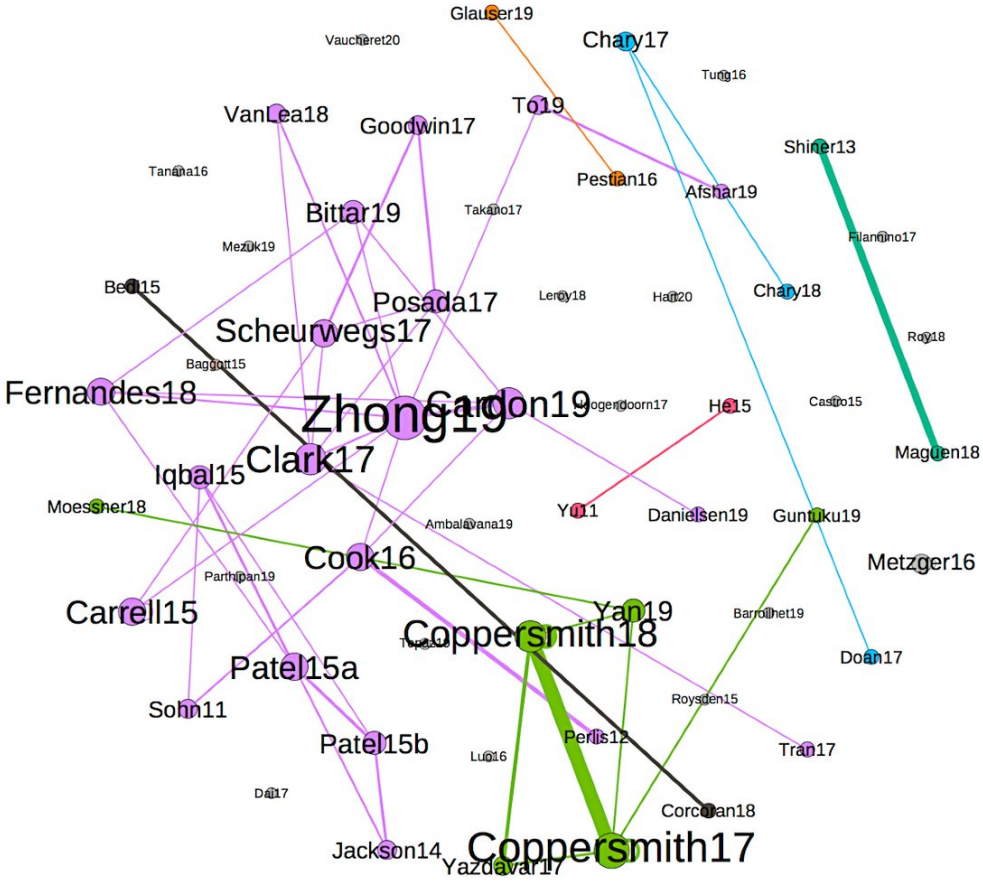


Figura 5. Grafico delle citazioni. Visualizza questa figura

4. DISCUSSIONE

4.1 PUNTI DI FORZA E LIMITI DELLA REVISIONE

Questo studio esamina i modelli ML e NLP nel campo della salute mentale, che è stato un tema di grande attualità negli

ultimi anni. La metodologia è stata elaborata per vagliare un numero massimo di studi medici specifici, estendendo la ricerca a 4 database medici (PubMed, Scopus, ScienceDirect e PsycINFO). Inoltre gli studi selezionati sono stati scelti con un occhio alla precisione e alla qualità, in modo da poter descrivere allo stesso tempo le popolazioni, i metodi, le fonti dei dati e gli aspetti tecnici.

Il limite principale di questo studio è la mancanza di confronti quantitativi tra gli studi selezionati. Non è infatti possibile confrontare studi altamente eterogenei senza modelli di ricerca comuni. Inoltre, i lavori selezionati non sono stati valutati in base al rischio di distorsione. Nonostante questa carenza, i loro limiti e punti di forza sono illustrati nelle singole tabelle dell'Appendice Multimediale 2.

4.1.1 Limiti metodologici e tecnici degli studi selezionati

I metodi ML e NLP possono essere considerati un nuovo paradigma nella ricerca medica, in cui diventa pratico analizzare ogni possibile e innovativo parametro di un argomento per discernere nuovi modelli clinici, anche se è inedito. Questo nuovo paradigma implica un ripensamento della metodologia standard, che consiste nel formulare un'ipotesi solida, definire gli obiettivi e raccogliere i risultati per confermare o respingere l'ipotesi. Tuttavia, nella pratica, gli studi selezionati tendono a confermare ipotesi cliniche basate su intuizioni cliniche fondamentali, ovvero le anomalie del linguaggio negli adulti con DSA [42].

Sono state rilevate altre limitazioni metodologiche e potenziali fonti di distorsione. Come indicato nella sezione Risultati, una delle 3 principali categorie di popolazione è costituita dagli *utenti di social network o chat* [40,41,66,77,79], i cui membri sono prevalentemente giovani. Per questo motivo, Coppersmith et al. [76,77] hanno avvertito che questi risultati potrebbero non essere generalizzabili ad altre popolazioni [77,106]. Inoltre, quando Chary et al [66] si sono concentrati sugli *utenti del Lycaeum* e Coppersmith et al [76] hanno citato i *partecipanti da un'azienda*, la

mancanza di informazioni precise sui partecipanti di una coorte era evidente. Fa eccezione il gruppo di utenti di OurDataHelps.org [77] che si sono offerti di partecipare a una ricerca scientifica e hanno compilato un questionario per fornire informazioni su di sé. Anche quando i partecipanti forniscono volontariamente informazioni personali, è molto probabile che le distorsioni di personalità giochino un ruolo, soprattutto negli studi sul suicidio e sulla depressione.

Allo stesso modo, gli studi raramente considerano le differenze culturali o etniche all'interno di un campione [80]. Ad esempio, in uno studio sul comportamento violento, i ricercatori dovrebbero riconoscere che le *sculacciate ai bambini a scopo disciplinare* sono considerate inappropriate in alcune culture, ma appropriate in altre. In alcuni casi, le caratteristiche specifiche della lingua possono migliorare le prestazioni dei metodi PNL. Ad esempio, nel caso di Takano et al. [62], la distribuzione dei morfemi viene utilizzata per distinguere tra ricordi specifici e non specifici nel test di memoria autobiografica. Come mostrato nel documento, tra i fattori distintivi più importanti vi sono le particelle grammaticali specifiche della lingua giapponese, come た/だ (tempo passato), ない (negazione), は (marcatore di argomento) e で (luogo o metodo). In lingue con strutture diverse, lo stesso metodo può essere meno efficace e può essere necessario studiare altri indicatori.

Esiste un vantaggio nell'utilizzo di ML e NLP per la pratica clinica della salute mentale?

Il principio caratteristico della ML è quello di analizzare grandi quantità di dati; tuttavia, questo a volte porta i ricercatori ad assumere implicitamente che *più dati vengono inseriti, più accurati saranno i risultati*. ML e NL consentono l'analisi di grandi quantità di dati e il confronto di ampi gruppi e pazienti. Per esempio, Roysden et al [56] hanno analizzato i dati amministrativi e gli EHR di una popolazione di 12.759 pazienti; Maguen et al [63] hanno confrontato oltre 8.168.330 note cliniche raccolte nell'arco di 15 anni; e Yazdavar et al [79] hanno analizzato i post scritti da 4.000 utenti di Twitter. Allo stesso tempo, anche se migliaia

di articoli sono stati pubblicati utilizzando dati medici, pochissimi hanno apportato dei contributi significativi alle pratiche cliniche [111].

Twitter e altri social network, con quasi 3 miliardi di utenti a livello globale, sono diventati fonti significative di informazioni per uso e ricerca medica [112]. Inoltre, l'analisi delle piattaforme basate sui social media può generare dettagli preziosi sulla salute mentale delle persone e sulle interazioni sociali o professionali. L'alterazione delle abitudini quotidiane è uno dei criteri fondamentali per la diagnosi di un disturbo della salute mentale (in generale, il criterio B del DSM-5). Un recente studio di Fagherazzi e Ravaud [113] illustra l'idea che l'IA possa essere implementata nel cosiddetto *digitosoma* (dati generati online e dalle tecnologie digitali) che costituisce un potente agente per individuare nuovi marcatori digitali e fattori di rischio in medicina. Analizzando una coorte globale di oltre 85.000 tweet settimanali scritti da persone affette da diabete, sono stati in grado di discutere i diversi modelli di stress legati alla malattia dei pazienti con diabete di tipo 1 o di tipo 2. Analizzando i tweet, Mowery et al. [106] hanno scoperto che possono esistere modi alternativi in cui le persone esprimono la depressione. Questi risultati indicano che potrebbero esserci degli altri modi per esprimere la malattia mentale.

Da questo punto di vista, le diverse espressioni del disagio psicologico (sia che le persone si rivolgano a operatori sanitari, a parenti o a reti di amicizie digitali) potrebbero essere accessibili e utili ai fornitori di cure. Il ML e l'NLP possono essere utili in psichiatria per identificare persone con rischi clinici di depressione, tentativi di suicidio, ansia o addirittura psicosi sulla base di dati digitali o note cliniche.

4.1.2 Riflessioni etiche

L'IA in psichiatria e più in generale in medicina solleva questioni etiche e richiede prudenza nella sua applicazione. Come accennato in precedenza, le tecniche di ML e NLP presentano vantaggi preziosi in psichiatria per l'analisi di grandi quantità

di dati con un'elevata validità diagnostica e prognostica. Questi strumenti, che hanno fatto scuola in medicina e psichiatria, dovrebbero ricevere maggiore attenzione per i loro promettenti risultati nella pratica clinica e nella ricerca medica. Inoltre, alcuni studi recenti suggeriscono che le persone si sentono più a loro agio quando parlano con una macchina rispetto a un medico: Lucas et al. [51] affermano che in uno studio clinico le persone che (credevano di) interagire con un computer hanno rivelato informazioni più apertamente di quelle che pensavano che fosse un individuo a controllare il computer. Forse la macchina è vista come più obiettiva di un essere umano e quindi riduce la paura del giudizio da parte del medico. L'introduzione di un computer come nuovo tipo di medico porta a un profondo cambiamento nel rapporto medico-paziente e promuove l'idea di un nuovo modello clinico che coinvolge un terzo individuo. Questa relazione è cruciale per la pratica clinica psichiatrica e l'uso dell'elaborazione dei dati dovrebbe essere discusso. Sassolas [114] ha messo in dubbio che la *psichiatria tecnologica* sia una pratica in grado di evitare quella che ha definito la "prossimità psichica della privacy". La psichiatria tecnologica potrebbe generare un incontro operativo il cui unico scopo è quello di normalizzare i sintomi del paziente e di ridurre la paura che questi vengano divulgati.

Oltre a migliorare le relazioni, l'applicazione della ML e della PNL in psichiatria dovrebbe avvenire con particolari precauzioni per evitare abusi clinici. Questa rassegna comprende due studi sulla previsione della psicosi in pazienti ad alto rischio di questa malattia. Uno di essi ha addirittura introdotto un modello di ML+NLP che ha mostrato un'accuratezza del 100% nel predire la psicosi in quest'ultimo campione di pazienti [39], migliore di una semplice valutazione clinica. Tuttavia, questi risultati devono essere trattati con cautela a causa delle piccole dimensioni del campione e della mancanza di dettagli sulle tecniche statistiche utilizzate. È necessario considerare il rischio di overfitting. Sebbene sia necessario proseguire la ricerca per migliorare gli aspetti tecnici, è necessario tenere conto dell'etica. Martinez-Martin et al. [115] si sono chiesti se sia etico utilizzare le stime prognostiche

del metodo ML per il trattamento della psicosi, in quanto non è noto se nel contesto locale siano presenti variabili (come le differenze nella pratica psichiatrica e nel supporto sociale) che potrebbero influenzare la validità del modello. Inoltre, quando si programma un algoritmo di ML, gli sperimentatori possono scegliere di rafforzare i criteri che ritengono più rilevanti, ad esempio i criteri clinici invece dei fattori socioeconomici. Ciò potrebbe essere meno adatto per alcuni pazienti quando l'analisi automatica della macchina dà l'illusione di una maggiore obiettività. Questi aggiustamenti dovrebbero essere fatti per rispettare il principio di equità.

Nel caso della previsione della psicosi, lo studio ha coinvolto solo pazienti che hanno acconsentito sia alle cure psichiatriche sia al completamento delle interviste. Non è stato così negli studi sulla prevenzione del suicidio, dove i ricercatori hanno raccolto informazioni sui pazienti utilizzando i social media. Questo potrebbe essere considerato una violazione della privacy. Le informazioni provenienti dai social media dovrebbero essere utilizzate per identificare i sintomi? L'applicazione dell'IA in questo contesto solleva notevoli preoccupazioni etiche, in particolare per quanto riguarda il bilanciamento tra buoni propositi e rispetto della riservatezza [53]. L'ML e l'NLP possono aiutare a identificare le persone a rischio clinico di depressione o di pensieri suicidi, che molto probabilmente non hanno accesso a operatori di salute mentale e/o a un medico di base [61]; tuttavia, questo riduce la protezione della privacy e può portare a una maggiore vulnerabilità in alcune popolazioni [21]. Per ottenere il consenso informato dei pazienti e proteggere la loro privacy, McKernan et al. [53] hanno raccomandato che i pazienti dovrebbero essere informati che (1) gli algoritmi possono essere imperfetti o sbagliati; (2) i dati degli algoritmi dovrebbero essere considerati altamente sensibili o di natura confidenziale; (3) i dati degli algoritmi potrebbero raccomandare azioni che non sono immediatamente evidenti; e (4) gli algoritmi potrebbero richiedere un intervento che in realtà non è necessario. Pertanto, gli psichiatri dovrebbero essere formati sulle tecniche di ML e NLP ed essere in grado

di spiegare ai pazienti le loro caratteristiche principali e il motivo per cui potrebbero richiedere determinate raccomandazioni. Quest'ultimo punto sottolinea la necessità di un'IA spiegabile che vada oltre i metodi a scatola nera.

Infine, la ML e la PNL non dovrebbero portare all'esautorazione degli psichiatri o sostituire il binomio medico-paziente. Al contrario, la combinazione di ML e PNL dovrebbe essere considerata uno *strumento di* supporto alla pratica clinica e alla ricerca medica.

5. CONCLUSIONI

Nell'ultimo decennio, l'uso di ML e NLP è diventato sempre più diffuso in medicina e più specificamente in psichiatria. Per questo motivo, la presente revisione si proponeva di riassumere e caratterizzare gli studi che hanno utilizzato tecniche di ML e NLP per la salute mentale in termini metodologici e tecnici. L'obiettivo secondario era quello di considerare l'uso potenziale di questi metodi nella pratica clinica della salute mentale (ad esempio, il contributo alla diagnosi, alla prognosi, alla definizione dei fattori di rischio, all'impatto della psicoterapia, all'aderenza al trattamento e agli effetti collaterali).

Sebbene gli studi selezionati fossero eterogenei in termini di argomenti e disturbi mentali, sono state riscontrate caratteristiche comuni in termini di categorie di popolazione (pazienti inclusi nei database medici, pazienti che si sono presentati al pronto soccorso e utenti dei social network) e obiettivi (ad esempio, estrazione dei sintomi, classificazione della gravità, confronto delle terapie, scoperta di indizi psicopatologici e correzioni alla nosografia attuale). L'analisi del tipo di dati utilizzati ha identificato due corpora principali: i dati raccolti dagli operatori sanitari (EHR, note cliniche o EMR) e i dati provenienti dai social media. Infine, l'analisi del metodo indica che gli autori hanno privilegiato alcune tecniche. I metodi standard di NLP (come la lemmatizzazione, il POS tagging o gli n-grammi) sono i più uti-

lizzati per la preelaborazione, oltre all'estrazione di CUI dedicata ai testi medici. L'analisi di classificazione specifica che si preferiscono i classificatori con buone prestazioni (SVM, LogIR e RF) rispetto a quelli con funzionamento *trasparente*. È stato verificato l'uso di piattaforme di linguaggio di programmazione universali come Python e R; Python è utilizzato più frequentemente e in tempi più recenti. L'analisi delle corrispondenze tra dati, metodi, classificatori, piattaforme e pubblicazioni rivela un gruppo di pubblicazioni che associano i dati delle note cliniche ai metodi cTAKES e alla piattaforma R-Python.

I metodi ML e NLP possono talvolta impressionare per l'enorme quantità di dati considerati e per le molteplici prospettive che offrono. Ciò ha portato alcuni autori a considerarli un nuovo paradigma nella ricerca sulla salute mentale. Tuttavia, questi processi tendono a confermare le ipotesi cliniche piuttosto che a sviluppare nuove informazioni, e alcuni risultati dovrebbero essere trattati con cautela (ad esempio, i risultati delle coorti di utenti di social media o l'impatto delle caratteristiche specifiche della lingua sulle prestazioni dei metodi NLP). Al contrario, le tecniche di ML e NLP forniscono informazioni su dati inesplorati e sulle abitudini quotidiane dei pazienti, solitamente inaccessibili agli operatori sanitari. Possono essere considerate uno strumento aggiuntivo in ogni fase della cura della salute mentale: diagnosi, prognosi, efficacia del trattamento e monitoraggio. A questo proposito, rimangono questioni etiche, come la previsione di disturbi psichiatrici o le implicazioni nel rapporto medico-paziente, che dovrebbero essere discusse in modo tempestivo. Pertanto, i metodi ML e NLP possono offrire molteplici prospettive nella ricerca sulla salute mentale, ma dovrebbero essere considerati come uno strumento di supporto alla pratica clinica.

LE FONTI DI QUESTO NUMERO

1. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022 Jul;79:102470. doi: 10.1016/j.media.2022.102470.
Tratto da: MEDICAL IMAGE ANALYSIS
Offre un forum per la diffusione dei nuovi risultati della ricerca nel campo dell'analisi delle immagini mediche e biologiche, con particolare attenzione agli sforzi relativi alle applicazioni della computer vision, della realtà virtuale e della robotica ai problemi di imaging biomedico. La rivista è interessata ad approcci che utilizzano set di dati di immagini biomediche a tutte le scale spaziali, dall'imaging molecolare/cellulare all'imaging di tessuti/organi.

2. Luchini C, Pea A, Scarpa A. Artificial intelligence in oncology: current applications and future perspectives. Br J Cancer. 2022 Jan;126(1):4-9. doi: 10.1038/s41416-021-01633-1.
Tratto da: BRITISH JOURNAL OF CANCER
È una delle riviste oncologiche generali più citate e si impegna a pubblicare ricerche all'avanguardia, traslazionali e cliniche sul cancro. La rivista accoglie ricerche su tutti i tipi di cancro e si concentra su: metastasi, microambiente, immunologia e immunoterapia, terapie mirate e di nuova generazione, chemioterapia e radioterapia, meccanismi di resistenza, studi clinici, genomica, epigenomica e medicina di precisione, epidemiologia, metabolismo e approcci diagnostici all'avanguardia.

3. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, Marsh J, DeVylder J, Walter M, Berrouiguet S, Lemey C. Machine Learning and Natural Language Processing in Men-

tal Health: Systematic Review. J Med Internet Res. 2021 May 4;23(5):e15708. doi: 10.2196/15708.
Tratto da: JOURNAL OF MEDICAL INTERNET RESEARCH
È una rivista leader nel campo dell'informatica sanitaria e dei servizi sanitari/politica sanitaria che si occupa di salute digitale, scienza dei dati, informatica sanitaria e tecnologie emergenti per la salute, la medicina e la ricerca biomedica.

4. Martinho A, Kroesen M, Chorus C. A healthy debate: Exploring the views of medical doctors on the ethics of artificial intelligence. Artif Intell Med. 2021 Nov;121:102190. doi: 10.1016/j.artmed.2021.102190.
Tratto da: ARTIFICIAL INTELLIGENCE IN MEDICINE
Pubblica articoli originali provenienti da un'ampia varietà di prospettive interdisciplinari riguardanti la teoria e la pratica dell'intelligenza artificiale (AI) in medicina, nella biologia umana ad orientamento medico e nell'assistenza sanitaria.

5. Abdelhalim H, Berber A, Lodi M, Jain R, Nair A, Pappu A, Patel K, Venkat V, D'Antoni F, Russo F, Ambrosio L, Vollero L, Vadalà G, Merone M, Papalia R, Denaro V. Artificial Intelligence and Computer Vision in Low Back Pain: A Systematic Review. Int J Environ Res Public Health. 2021 Oct 17;18(20):10909. doi: 10.3390/ijerph182010909.
Tratto da: INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH AND PUBLIC HEALTH
È una rivista scientifica peer-reviewed che pubblica articoli originali, recensioni critiche, note di ricerca e brevi comunicazioni nell'area interdisciplinare delle scienze della salute ambientale e della salute pubblica. Collega diverse discipline scientifiche, tra cui biologia, biochimica, biotecnologia, biologia cellulare e molecolare, chimica, informatica, ecologia, ingegneria, epidemiologia, genetica, immunologia, microbiologia, oncologia, patologia, farmacologia e tossicologia, in modo integrato, per affrontare le questioni critiche legate alla qualità ambientale e alla salute pubblica.

6. Manickam P, Mariappan SA, Murugesan SM, Hansda S, Kaushik A, Shinde R, Thipperudraswamy SP. Artificial Intelligence (AI) and Internet of Medical Things (IoMT) Assisted Biomedical Systems for Intelligent Healthcare. Biosensors (Basel). 2022 Jul 25;12(8):562. doi: 10.3390/bios12080562.
Tratto da: BIOSENSORS
Offre un forum avanzato per gli studi relativi alla scienza e alla tecnologia dei biosensori e del biosensing. Pubblica articoli di ricerca originali, recensioni complete e comunicazioni.

7. Zia A, Aziz M, Popa I, Khan SA, Hamedani AF, Asif AR. Artificial Intelligence-Based Medical Data Mining. J Pers Med. 2022 Aug 24;12(9):1359. doi: 10.3390/jpm12091359. tratto da:
Tratto da: JOURNAL OF PERSONALIZED MEDICINE
È una rivista internazionale ad accesso libero che mira a riunire tutti gli aspetti della medicina personalizzata in un'unica piattaforma. JPM pubblica ricerche scientifiche innovative e all'avanguardia, pre cliniche e traslazionali, e tecnologie relative alla medicina personalizzata (ad esempio, medicina di precisione, farmaco genomica/proteomica, biologia dei sistemi, analisi di associazione 'omica').

8. Zhang Y, Weng Y, Lund J. Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. Diagnostics (Basel). 2022 Jan 19;12(2):237. doi: 10.3390/diagnostics12020237.
Tratto da: DIAGNOSTICS
È una rivista scientifica internazionale ad accesso libero sulla diagnostica medica. I campi di ricerca di interesse includono, tra gli altri, radiologia, medicina nucleare, imaging medico, endoscopia, patologia e diagnostica molecolare.

9. Bhardwaj A, Kishore S, Pandey DK. Artificial Intelligence in Biological Sciences. Life (Basel). 2022 Sep 14;12(9):1430. doi: 10.3390/life12091430.
Tratto da: LIFE
È una rivista scientifica internazionale, peer-reviewed e ad accesso aperto, che tratta temi fondamentali delle scienze della vita, dalla ricerca di base a quella applicata. I campi di ricerca di interesse includono, tra gli altri, vita artificiale, astrobiologia, biochimica, biodiversità, bioinformatica e biomedicina.

10. Korteling JEH, van de Boer-Visschedijk GC, Blankendaal RAM, Boonekamp RC, Eikelboom AR. Human- versus Artificial Intelligence. Front Artif Intell. 2021 Mar 25;4:622364. doi: 10.3389/frai.2021.622364.
Tratto da: FRONTIERS IN ARTIFICIAL INTELLIGENCE
Frontiers in Artificial Intelligence pubblica ricerche rigorosamente peer-reviewed sulla rivoluzione tecnologica dirompente e all'avanguardia dell'Intelligenza Artificiale (IA). Si propone di essere all'avanguardia nella diffusione delle conoscenze scientifiche e delle scoperte d'impatto per gli accademici, i responsabili politici, l'industria e il pubblico di tutto il mondo. Il modello ad accesso aperto e la rigorosa e rapida revisione tra pari fanno della rivista la principale piattaforma per la pubblicazione di contenuti di alta qualità sulle sfide e le soluzioni che portano la suite di metodi dell'intelligenza artificiale all'uso pratico.