Niccolò Puccinelli, Matricola 881395

CASO AMA

"È etico utilizzare stime prognostiche elaborate con il machine learning per trattare forme di psicosi?"

INTRODUZIONE

Il dott.K (psichiatra) e i suoi colleghi stanno valutando l'idea di includere, all'interno della loro pratica clinica relativa al trattamento di pazienti psicotici, un modello predittivo addestrato su un dataset multisito europeo. In particolare, grazie al Machine Learning (ML), tale modello è in grado di stimare, secondo le statistiche e previa conoscenza di alcuni dati personali, la condizione psicotica del paziente a un anno dal primo episodio di psicosi. Una volta fornita la stima prognostica, l'algoritmo di ML propone, per alcuni pazienti, determinati trattamenti terapeutici.

Spesso i pazienti afflitti da psicosi non guariscono mai del tutto, o perché non hanno ricevuto cure, o perché il trattamento è stato iniziato troppo tardi[1]. Per questo particolare ambito l'innovazione clinica si pone dunque come necessaria per garantire un aumento sostanziale del tasso di guarigione. Ciò nonostante, il modello proposto dal dott.K presenta numerosi problemi, di carattere sia tecnico, sia (soprattutto) etico, rendendo nebulosa la questione dell'accettabilità etica di questo metodo.

PROBLEMI TECNICI

I principali problemi tecnici derivano dall'impiego di un modello di ML e dalla scelta di utilizzare un dataset che, seppur grande, non può essere completamente rappresentativo degli specifici pazienti in cura dal dott.K. L'algoritmo predittivo è infatti stato addestrato esclusivamente sul database multisito europeo e non ha ricevuto modifiche e/o correzioni, da effettuarsi tenendo conto dello specifico contesto ospedaliero in cui lavorano il dott.K e i suoi colleghi. Alla luce della diversità e della specificità dei pazienti in cura alla clinica ospedaliera, il modello potrebbe vedere ridotta la propria efficacia terapeutica. Tale algoritmo si basa infatti su una logica *subject-wise*, i.e. le fasi di *training* e di *testing* statistico vengono eseguite su soggetti diversi rispetto alla fase di applicazione clinico-pratica (a differenza della logica *record-wise*, in cui il modello viene addestrato, testato e applicato sugli stessi soggetti)[2]. Poiché ogni paziente possiede caratteristiche uniche circa i propri parametri diagnostici e considerando l'eterogeneità di sintomi e cause della psicosi, l'impiego di un algoritmo *subject-wise* per predire le condizioni psicotiche dei pazienti registrati nel database non garantisce la medesima accuratezza sugli assistiti del dott.K.

Inoltre, l'utilizzo di un database e, più in generale, dei *big data*, implica, specialmente in campo medico, l'insorgere di problemi di *bias* (come quello precedentemente descritto riguardo all'eccessivo adattamento ai parametri dei pazienti presenti nell'archivio europeo) e di input, ovvero dati poco precisi ugualmente processati ed elaborati dal modello. L'accuratezza del modello (75%) è quindi inesorabilmente soggetta ad un calo. Questo significa che a più di un paziente su quattro la macchina suggerirà una scelta terapeutica non ideale, se non addirittura completamente sbagliata. Può succedere, ad esempio, che il modello suggerisca l'impiego di farmaci antipsicotici anche a soggetti non psicotici. Per di più, la validità dei dati è stata testata esclusivamente a livello statistico. Clinicamente, non è stata ancora accertata la superiorità dell'algoritmo rispetto alle tecniche tradizionali per la cura dei pazienti affetti da psicosi.

Nonostante sia il medico ad occuparsi della proposta di trattamento da fare al paziente (il modello si occupa "solo" di suggerire la terapia da applicare), il modo in cui l'algoritmo usa i dati di input per giungere ad una

[1] https://www.therecoveryvillage.com/mental-health/psychosis/psychosis-statistics/

[2] Max A Little, Gael Varoquaux, Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C Mohr, Konrad P Kording, Using and understanding cross-validation strategies. Perspectives on Saeb et al., *GigaScience*, Volume 6, Issue 5, May 2017

decisione in merito potrebbe risultare poco trasparente (modello *black box*)[3,4]. Dal punto di vista tecnico, questo costituirebbe un problema qualora il dott.K decidesse di apportare una serie di modifiche e/o correzioni, includendo i casi di psicosi relativi ai pazienti della propria clinica. Inoltre, il dottore e gli altri membri dello staff medico che utilizzano il modello predittivo hanno bisogno di essere istruiti sul preciso funzionamento dello stesso, sia per comprenderlo, sia per riuscire a comunicarne il funzionamento in maniera chiara. Ciò risulta difficoltoso a causa della natura generalmente opaca delle tecniche di ML.

Bisogna inoltre considerare l'ubicazione della clinica del dott.K. Nell'eventualità in cui questa fosse (ad esempio) americana e offrisse i propri servizi al di fuori dell'UE, sorgerebbero altre criticità riguardanti il trattamento dei dati personali (in assenza del GDPR dell'Unione Europea) e l'impiego del dataset europeo. Quest'ultimo infatti potrebbe contenere variabili che, extra-UE, invaliderebbero il modello. Tuttavia, di seguito sarà considerato il caso in cui la clinica si trovi in territorio UE.

Infine, il crescente affidamento all'IA da parte del dott.K e della sua *équipe* può inficiare, più o meno gravemente, il loro giudizio. Nel breve periodo ciò si potrebbe tradurre in *overreliance*, i.e. nel crescente affidamento alle capacità decisionali del modello di ML. Nel lungo periodo invece, l'eccessiva automatizzazione offerta dall'IA potrebbe portare ad una vera e propria perdita di abilità (*deskilling*) da parte della *troupe* ospedaliera[4].

Molti di questi problemi tecnici nascondono diverse problematiche di tipo etico, alle quali si aggiungono le questioni relative al consenso del paziente, alla privacy dei dati personali, alla comunicazione.

PROBLEMI ETICI

In primo luogo, è essenziale una riflessione sull'utilizzo di un database multisito europeo. Sebbene il vantaggio di uno studio effettuato su basi di dati diverse sia proprio quello di offrire prospettive cliniche differenti, queste sono tuttavia legate dai medesimi codici dei pazienti che, seppur anonimi, forniscono vari dati sensibili riferiti ad individui specifici. Esiste dunque, a monte, un problema di privacy riguardante gli individui oggetto di studio per l'addestramento del modello di ML. L'unione definitiva di tali archivi in un'unica base di dati potrebbe far perdere i riferimenti necessari per tenere traccia dei soggetti, ma ciò comporterebbe la perdita della pluralità di prospettive e il pericolo di monopolio dei dati relativi alla salute da parte di un'unica organizzazione.

Esaminando il modello di ML *per sé*, i problemi principali derivano dalla struttura *black box*. Anzitutto, la natura opaca con cui opera questa tipologia di modelli risulta imperscrutabile non solo ai medici, ma spesso persino agli stessi sviluppatori, i quali implementano tecniche di ML in grado di restituire un determinato output senza che i procedimenti compiuti durante il calcolo risultino espliciti[4]. Questo porta inevitabilmente anche ad un problema di responsabilità. Non possedendo abbastanza informazioni sul modo in cui opera il modello, il rischio, in caso di calcoli erronei, è quello di un continuo "rimbalzo" di responsabilità tra medici, clinica e sviluppatori/ingegneri del software. Oltretutto, un algoritmo *black box* non garantisce sicurezze riguardo la finalità etica di chi l'ha ideato o progettato. Il modello potrebbe infatti indurre ad un consumo di farmaci maggiore per scopi puramente economici, dando luogo ad una discrepanza di obiettivi tra chi lo impiega per profitto personale e chi lo applica esclusivamente in ottica medica.
Il Considerando 71 del *General Data Protection Regulation* (GDPR) dell'Unione Europea, entrato in vigore nel 2018, recita: "[…] *In ogni caso, tale trattamento dovrebbe essere subordinato a garanzie adeguate, che dovrebbero comprendere la specifica informazione all'interessato e il diritto di ottenere l'intervento umano, di esprimere la propria opinione, di ottenere una spiegazione della decisione conseguita dopo tale valutazione e di contestare la decisione.* […]". Pertanto, il dott.K e i suoi colleghi dovrebbero essere in grado di capire il funzionamento del modello in maniera appropriata, affinché possano partecipare attivamente ai processi decisionali e fornire una spiegazione esplicita al paziente.

[3] N. Musacchio, G. Guaita, A. Ozzello, M.A. Pellegrini, P. Ponzani, R. Zilich, A. De Micheli (2018) Intelligenza Artificiale e Big Data in ambito medico: prospettive, opportunità, criticità. JAMD Vol. 21-3

[4] Federico Cabitza et al., Unintended Consequences of Machine Learning in Medicine, Jama, 2017, 318:6

Il dott.K potrebbe dunque rivolgersi direttamente ai programmatori e agli ingegneri del software, cercando di ottenere (e comprendere) quante più informazioni possibili sul modo in cui opera l'algoritmo. In alternativa, i medici potrebbero utilizzare l'IA solo come "partner" per eseguire la stima prognostica, i.e. senza la predominanza di quest'ultima nel processo decisionale, ma esclusivamente per eventuale conferma dello stato psicotico dei soggetti, evitando quindi del tutto una somministrazione di farmaci suggerita da un metodo poco trasparente, che peraltro non tiene conto dei dati specifici di ogni paziente.

Numerosi problemi etici sono inoltre sollevati dal preminente utilizzo dell'IA che tale tecnica comporta. Il modello infatti può potenzialmente occuparsi sia della diagnosi che della terapia dei pazienti psicotici, concedendo al dott.K un mero ruolo tecnico-interpretativo dei risultati prodotti dalla macchina. Tale progressivo (ed eccessivo) affidamento al metodo di ML può portare ai due fenomeni descritti precedentemente nella parte concernente i problemi tecnici (*overreliance* e *deskilling*), ma non solo: lo smisurato impiego dell'IA va di pari passo con il venire meno del rapporto olistico con l'assistito, mediante un approccio sempre meno attento agli aspetti psicologici e sociali (datificazione[4]).
In secondo luogo, l'alta invasività decisionale del modello di ML, in ambito sia prognostico che terapeutico, pone un ulteriore quesito etico fondamentale: cosa fare se i pareri di medici e macchina sono discordanti? Anche se l'accuratezza statistica del modello (75%) rappresenta un notevole miglioramento rispetto all'approccio tradizionale, i medici della clinica sono in possesso di informazioni specifiche per ogni soggetto, precluse all'algoritmo.
Al fine di mitigare l'analisi decisionale del modello, potrebbe essere necessaria l'introduzione di un *human-in-the-loop*[5] con capacità di *override* del sistema. In questo caso, il dott.K dovrebbe prima "aprire la scatola nera" dell'algoritmo per comprenderne il funzionamento, per poi inserirsi all'interno del processo decisionale, controllandone i passaggi e, soprattutto, assicurandosi (preferibilmente insieme alla sua *équipe* medica) di avere sempre l'ultima parola sulla diagnosi e sul successivo trattamento della psicosi.
Un'altra possibile strada percorribile, che tuttavia non esclude la precedente, prevede l'adattamento del modello agli specifici casi clinici in cura dal dott.K. L'algoritmo potrebbe infatti essere riaddestrato e calibrato in conformità ai soggetti della clinica. Ciò garantirebbe un aumento dell'accuratezza predittiva e una minore genericità della terapia, senza tuttavia colmare il vuoto relativo all'approccio olistico al paziente, del quale dovrebbero comunque occuparsi i medici in prima persona.
Eppure, questa possibile soluzione apre le porte ad un ulteriore problema, riguardante la privacy.

Un qualsiasi modello di ML è intrinsecamente caratterizzato dall'apprendimento automatico delle informazioni. Supponendo che in fase di applicazione pratica il modello impieghi solo parzialmente le informazioni cliniche del paziente per automigliorarsi, se il dott.K decidesse di riaddestrarlo usando come *training-set* i dati dei pazienti della clinica, ad un aumento della precisione corrisponderebbe una maggiore violazione della privacy. Gli stessi dati che il modello ha bisogno di conoscere per effettuare una predizione sullo stato psicotico, quali le storie relative ad episodi depressivi precedenti, sono strettamente personali. Perciò, addirittura addestrare l'algoritmo a partire da suddette informazioni significherebbe, in un certo senso, mantenerle in memoria. Nondimeno, la struttura *black box* del modello non garantisce sufficiente trasparenza e non vi è alcuna sicurezza sul modo in cui le informazioni siano trattate e utilizzate dalla macchina. Il paziente dovrà quindi essere informato adeguatamente, in modo chiaro, comprensibile e completo (rispetto ad ogni caso specifico), dal dott.K e dai suoi colleghi, in merito ai vantaggi e agli svantaggi correlati all'impiego di questo modello predittivo.

Di conseguenza, si può rilevare un ultimo evento critico nella comunicazione al paziente. Il dott.K ha infatti a che fare con diversi ostacoli comunicativi, concernenti la struttura *black box* del modello (che non garantisce chiarezza né ai medici, né tantomeno ai pazienti), l'utilizzo dei *big data* (con le relative problematiche quali validità, qualità e corretto utilizzo dei dati), la probabilità che l'algoritmo sbagli (tra l'altro il modello non è ancora stato clinicamente testato, sebbene raggiunga il 75% di accuratezza statistica), i rischi relativi alla privacy e la possibile difficoltà nel comunicare con pazienti psicotici[6]. La situazione in cui si trovano il dott.K e i suoi colleghi risulta assai complicata e dipende fortemente dall'"apertura della scatola nera", grazie alla quale è possibile prevedere tre scenari possibili.
Nel primo caso, il dott.K non riesce ad ottenere (*disclosure*) o comprendere (*understanding*) sufficienti

[5] Gómez-González, E. and Gómez, E., *Artificial Intelligence in Medicine and Healthcare: applications, availability and societal impact*, EUR 30197 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-18454-6, doi: 10.2760/047666, JRC120214

[6] Supporto etico in medicina, Accademia Svizzera delle Scienze Mediche (ASSM), Casa delle accademie, Laupenstrasse 7, CH-3001 Berna, Howald Fosco Biberstein, Basilea, 2017

informazioni sul modello, mentre, nel secondo caso, egli comprende il funzionamento delle tecniche di ML utilizzate, ma si accorge che vi sono rischi molto alti (ad esempio per la privacy) per i soggetti in cura. In entrambe le eventualità, i medici potrebbero scegliere di utilizzare solo una parte del modello (se questa non presenta fattori di rischio elevati), oppure scegliere di non adoperarlo del tutto. Nel terzo caso, il dott.K comprende pienamente la natura dell'algoritmo e individua fattori di rischio bassi o moderati, per cui potrebbe decidere di impiegare l'algoritmo nella sua totalità (con eventuale *deskilling* progressivo dell'*équipe* medica) o parzialmente.
Ad ogni modo, il dottore deve descrivere al paziente i rischi e i benefici derivanti dall'uso di questa tipologia di IA, fornendogli informazioni imparziali e specificatamente complete, assicurandosene la comprensione[5]. Le informazioni dovranno dunque includere i vantaggi (come l'elevata precisione statistica rispetto ai metodi tradizionali e la sperimentazione su un grande database multisito) e gli svantaggi sopracitati, i.e. grado di opacità del modello, rischi legati all'impiego dei *big data*, probabilità di errore (e relative conseguenze) e pericoli per la privacy. Inoltre, al paziente andranno comunicati ulteriori pro e contro legati a precise scelte prese dal dott.K e dai suoi colleghi, quali un eventuale adattamento del modello e/o un possibile approccio *human-in-the-loop*.
In aggiunta, il dott.K ha a che fare con soggetti psicotici e ciò rende ancora più ardua la comunicazione. A seconda del livello di psicosi, i pazienti potrebbero accettare (o rifiutare) il trattamento senza averlo opportunamente capito, non essere capaci di intendere e di volere o, in casi estremi, rivelarsi un pericolo per sé stessi e per gli altri. I medici potrebbero dunque scegliere di modulare ulteriormente l'informazione a seconda del paziente (senza tuttavia snaturarla), oppure, in casi gravi, contattare i parenti dell'assistito, ai quali dovranno essere spiegati vantaggi e svantaggi della terapia sviluppata tramite tecniche di ML. In sostanza, il paziente (o chi per lui) dovrà avere compreso chiaramente i rischi e i benefici derivanti dall'impiego dell'IA, in modo da fornire un consenso/dissenso pienamente informato.

CONCLUSIONI

In ultima analisi, per quanto riguarda la stima prognostica, la rivelazione di quest'ultima dovrebbe far parte del consenso informato sottoscritto dal paziente. L'articolo 4 del GDPR definisce i dati relativi alla salute come "*i dati personali attinenti alla salute fisica o mentale di una persona fisica* […] *che rivelano informazioni relative al suo stato di salute*" e il Considerando 63 del GDPR recita "*Un interessato dovrebbe avere il diritto di accedere ai dati personali raccolti che la riguardano e di esercitare tale diritto facilmente e a intervalli ragionevoli, per essere consapevole del trattamento e verificarne la liceità. Ciò include il diritto di accedere ai dati relativi alla salute, ad esempio le cartelle mediche contenenti informazioni quali diagnosi, risultati di esami, pareri di medici curanti o eventuali terapie o interventi praticati.* […]". Di conseguenza, la rivelazione della propria stima prognostica è un diritto del paziente. Inoltre, prendendo in esame il caso italiano, l'articolo 1, comma 3, della Legge 219/2017, afferma che "*Ogni persona ha il diritto di conoscere le proprie condizioni di salute e di essere informata in modo completo, aggiornato e a lei comprensibile riguardo alla diagnosi, alla prognosi, ai benefici e ai rischi degli accertamenti diagnostici e dei trattamenti sanitari indicati, nonché riguardo alle possibili alternative e alle conseguenze dell'eventuale rifiuto del trattamento sanitario e dell'accertamento diagnostico o della rinuncia ai medesimi.*".

Pertanto, il dott.K e la sua *équipe* sono tenuti ad informare adeguatamente il paziente (o chi per lui, in caso di psicosi altamente debilitante) riguardo alla possibilità di conoscere la propria stima prognostica (con annessa percentuale di accuratezza statistica) e, qualora egli ne manifesti la volontà, a comunicargliela in modo appropriato.