

UNIVERSITÀ  
DEGLI STUDI  
DI PADOVA



DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA BIOMEDICA

# **Rivelazione automatica e analisi di anomalie nei dati delle cartelle cliniche elettroniche**

**Relatore**

Prof. Giovanni Sparacino

**Laureando**

Alessio Canazza

ANNO ACCADEMICO 2024-2025

Data di laurea 25/09/2025



# Sommario

Negli ultimi decenni, il settore sanitario è stato protagonista di una profonda trasformazione grazie all'applicazione delle tecnologie digitali. Uno dei principali obiettivi della digitalizzazione del sistema sanitario è la progressiva sostituzione delle cartelle cliniche cartacee con le cartelle cliniche elettroniche o Electronic Health Records (EHR), strumento dalle molteplici potenzialità che può considerevolmente portare ad un miglioramento della qualità assistenziale del paziente e della ricerca in campo medico. Queste potenzialità sono tuttavia accompagnate da criticità relativamente all'affidabilità e alla sicurezza delle informazioni che vi sono contenute.

All'interno delle cartelle cliniche elettroniche si possono riscontrare delle anomalie nei dati che vi sono contenuti, intese come dati o parametri anomali e non coerenti rispetto alla realtà o alle aspettative che si possono formulare dal contesto e dalle conoscenze mediche. In particolare, queste anomalie possono essere dovute a un'errata archiviazione dei dati del paziente, a un accesso illecito alla cartella clinica o alla presenza di valori anomali nei parametri fisiologici del paziente. Spesso queste anomalie sono nascoste all'interno dei dati presenti nella cartella clinica e difficili da individuare; tuttavia, se trascurate, potrebbero mettere a rischio la sicurezza e la privacy del paziente, oltre che compromettere la qualità del servizio di assistenza sanitaria erogato.

Vista la notevole quantità di dati contenuti nei database sanitari e l'importanza di tali dati per la salute individuale e pubblica, risultano di fondamentale importanza sistemi di controllo per la ricerca automatica di anomalie.

Il presente elaborato si pone l'obiettivo di fornire una panoramica sulle attuali tecnologie in materia, andando ad approfondire le diverse tipologie di anomalie che si possono riscontrare in una cartella clinica elettronica e alcune delle tecnologie dedicate alla loro rilevazione.

Nel capitolo 1 verrà presentato lo strumento della cartella clinica elettronica e le sue principali caratteristiche; nel capitolo 2 verranno introdotte le diverse tipologie di anomalie che si possono riscontrare in una cartella clinica elettronica, ovvero le anomalie da dati implausibili (sez. 2.2), le anomalie da accessi illeciti (sez. 2.3) e le anomalie nei parametri del paziente (sez. 2.4). Il capitolo 3 tratterà due algoritmi per la rilevazione di anomalie determinate da dati implausibili: "FinFFPOF" (sez. 3.1) e "autoencoder" (sez. 3.2). Il capitolo 4 discuterà "Isolation Forest"

(sez. 4.1) e "Local Outlier Factor" (sez. 4.2), due algoritmi per la rilevazione delle anomalie determinate da accessi illeciti. Infine, il capitolo 5 tratterà le anomalie dovute a parametri critici o errori clinici, presentando l'algoritmo "Higher-Order Tensor Network" (sez. 5.1).

# Indice

<b>1</b>	<b>La cartella clinica elettronica: aspetti generali</b>	<b>1</b>
1.1	La cartella clinica . . . . .	1
1.2	La cartella clinica elettronica . . . . .	2
1.3	Vantaggi della cartella clinica elettronica . . . . .	3
1.3.1	Supporto fisico . . . . .	3
1.3.2	Accessibilità e sicurezza . . . . .	3
1.3.3	Organizzazione dati e ricerca . . . . .	3
1.3.4	Caratteristiche di attività . . . . .	4
1.4	Architettura dei sistemi delle cartelle cliniche elettroniche . . . . .	5
1.4.1	Struttura mainframe . . . . .	5
1.4.2	Struttura client-server . . . . .	6
1.5	Il Fascicolo Sanitario Elettronico . . . . .	7
<b>2</b>	<b>Possibili anomalie nelle cartelle cliniche elettroniche e rischi connessi</b>	<b>11</b>
2.1	La problematica . . . . .	11
2.2	Anomalie da dati implausibili . . . . .	12
2.2.1	Esempi . . . . .	12
2.2.2	Possibili cause . . . . .	13
2.2.3	Possibili strategie di riconoscimento automatico . . . . .	13
2.3	Anomalie da accessi illeciti . . . . .	14
2.3.1	Esempio . . . . .	15
2.4	Anomalie nei parametri del paziente . . . . .	16
2.4.1	Esempio . . . . .	16
2.5	Difficoltà attuali nella rivelazione automatica di anomalie . . . . .	17
<b>3</b>	<b>Rilevazione e analisi delle anomalie determinate da dati non plausibili</b>	<b>19</b>
3.1	Algoritmi tipo FindFPOF - Frequent Pattern Outlier Factor . . . . .	19
3.2	Algoritmi basati su autoencoder . . . . .	21

3.3	Una applicazione ai registri oncologici . . . . .	22
<b>4</b>	<b>Rilevazione e analisi delle anomalie determinate da accessi illeciti</b>	<b>27</b>
4.1	Isolation Forest . . . . .	27
4.2	Local Outlier Factor (LOF) . . . . .	29
4.3	Un'applicazione all'analisi di cartelle in un ospedale inglese . . . . .	31
4.3.1	Preparazione dei dati . . . . .	31
4.3.2	Metriche di valutazione . . . . .	32
4.3.3	Risultati . . . . .	32
<b>5</b>	<b>Rilevazione e analisi delle anomalie dovute a parametri critici o errori clinici</b>	<b>35</b>
5.1	Rilevazione delle anomalie tramite Higher-Order Tensor Network . . . . .	35
5.1.1	Rappresentazione delle sequenze di eventi nella rete tensionale multidimensionale . . . . .	36
5.1.2	Calcolo del punteggio di anomalia . . . . .	38
5.1.3	Risultati . . . . .	39
<b>6</b>	<b>Conclusioni</b>	<b>41</b>
	<b>Bibliografia</b>	<b>43</b>

# Capitolo 1

## La cartella clinica elettronica: aspetti generali

### 1.1 La cartella clinica

La cartella clinica [1][2] è il nucleo fondamentale dell'assistenza sanitaria, contenente tutte le informazioni e i dati clinici di un paziente raccolti durante gli incontri con gli operatori sanitari utili ai fini di prevenzione e di ricovero in caso di malattia. Una prima forma di cartella clinica la si può osservare già nel IV secolo a.C. quando Ippocrate, presso il tempio di Asclepio (Epidauro, Grecia), raccoglie dati come nome, città di appartenenza, diagnosi e terapia dei pazienti ricevuti. In Italia la cartella clinica diventa un documento obbligatorio a partire dal 1890.

La principale funzione della cartella clinica è quella di essere un documento informativo; essa infatti raccogliendo i dati anagrafici e la narrazione dello stato del paziente durante tutto il processo di cura, costituisce un importante strumento di comunicazione asincrona tra medici che porta ad una più tempestiva ed efficiente cura del paziente. Inoltre, la cartella clinica è un'importante documento di valore medico-legale. Infatti, in fase di dimissione, rappresenta l'atto ufficiale che testimonia tutti i trattamenti e le procedure eseguite sul paziente e risulta essere un elemento chiave per il calcolo delle tariffe di rimborso spettanti alla struttura sanitaria sede del ricovero secondo il modello dei DRG attualmente in vigore in Italia.

La cartella clinica nasce naturalmente come uno strumento cartaceo ma, grazie alle nuove tecnologie in ambito medico e tecnologico il flusso di informazioni cliniche da trasmettere, archiviare ed elaborare, ha di recente subito una crescita esponenziale che risulta essere gestibile solo con l'ausilio di calcolatori e reti telematiche, inserendo quindi questo strumento nel più ampio contesto di informatizzazione del sistema sanitario.

## 1.2 La cartella clinica elettronica

L'Unione Europea definisce la cartella clinica elettronica come una raccolta di dati sanitari elettronici relativi ad una persona fisica, rilevanti nell'ambito del sistema sanitario, il cui trattamento avviene ai fini dell'erogazione dell'assistenza sanitaria.[3]

In particolare invece, per la normativa internazionale ISO/TR 20514, l'Electronic Health Record (EHR) è un archivio di informazioni riguardanti la salute di un soggetto in cura, in un formato elaborabile da un computer, archiviato e trasmesso in modo sicuro e accessibile a più utenti autorizzati. Ha un modello informativo standardizzato, indipendente dal sistema EHR stesso, e il suo scopo è quello di supportare un'assistenza sanitaria continua, efficiente e di qualità attraverso il contenuto di informazioni retrospettive, simultanee e prospettive.[4]

L'archiviazione digitale delle cartelle cliniche diventa possibile in Italia a partire dal 2013 attraverso il D.L. 179/2012. Attualmente però, si considera digitalizzazione della cartella clinica anche la sola scansione del corrispondente documento cartaceo; tuttavia, in questo modo, non vengono sfruttati tutti i potenziali vantaggi che comporta l'utilizzo della versione digitale. Una cartella clinica cartacea infatti viene prodotta in corrispondenza di un determinato e singolo evento clinico del paziente, quale un ricovero, e su tale evento si concentra la sua narrazione. Una cartella clinica elettronica può invece essere considerata "*long life*" in quanto, in virtù della sua natura accessibile, può raccogliere i dati dell'intera storia clinica del paziente, permettendo un'assistenza sanitaria che non si concentra sul singolo evento clinico ma sul costante benessere del paziente, che inizia fin dallo stato di salute e continua durante la malattia o il ricovero e fino alla fase di recupero o di riabilitazione. Per questo motivo attualmente risulta essere uno dei principali obiettivi la completa ed efficiente transizione dalla cartella clinica cartacea alla cartella clinica elettronica nativa.[5]

Le funzionalità fondamentali che deve avere una cartella clinica elettronica sono le seguenti:

- **Archiviazione di dati sanitari:** deve essere in grado di memorizzare e conservare tutte le informazioni sullo stato di salute del paziente quali la sua storia clinica, le sue allergie, i referti degli esami, diagnosi e terapie. La raccolta di queste informazioni deve essere organizzata in modo tale da poter aiutare il medico a prendere decisioni cliniche appropriate, per un'assistenza sanitaria mirata.
- **Interazione efficiente:** la cartella clinica non deve essere un'ostacolo per l'operatore sanitario ma, al contrario, deve consentirgli di lavorare efficientemente e con risparmio di tempo sia in fase di scrittura o di raccolta dati che in fase di consultazione.

## **1.3 Vantaggi della cartella clinica elettronica**

Si elencano di seguito alcuni dei principali vantaggi relativi alla scelta di una cartella clinica elettronica rispetto alla versione cartacea.

### **1.3.1 Supporto fisico**

Le cartelle cliniche cartacee richiedono un supporto fisico spesso ingombrante e delicato, che può danneggiarsi nel tempo oltre che richiedere interi archivi per la conservazione. Le cartelle cliniche elettroniche invece, hanno il vantaggio di occupare molto meno spazio fisico: molti Gigabyte di dati possono essere conservati in pochi centimetri cubi di spazio, con un notevole risparmio di spazio fisico. La loro conservazione richiede comunque strutture che necessitano di specifiche caratteristiche, i server dati, ma consentono la creazione di copie di backup che ne impediscono la perdita accidentale. Hanno inoltre il vantaggio di poter essere conservate per un periodo di tempo potenzialmente infinito, o comunque definito solamente dai limiti di legge previsti in materia di privacy; questo potenziale permette di creare una cartella clinica del paziente duratura, comprendente dati raccolti durante tutta la vita del paziente, con risvolti positivi in ambito medico.

### **1.3.2 Accessibilità e sicurezza**

Costituisce uno degli aspetti fondamentali delle cartelle cliniche elettroniche. In primo luogo attraverso la creazione di sistemi di autenticazione e di controllo è possibile fornire le autorizzazioni di accesso ai dati solo al personale realmente necessario, rendendo possibile monitorare, attraverso i registri di log, ogni singolo accesso ai dati del paziente, controllando potenziali violazioni del sistema o accessi illeciti. Viene quindi potenziata la privacy dei dati contenuti nella cartella. Inoltre, la cartella in formato digitale permette una rapida diffusione e duplicazione, con la possibilità di trasportare informazioni per grandi distanze in breve tempo con costi che si possono considerare trascurabili. Infine, il supporto digitale rende possibile l'accesso ai dati che vi sono contenuti a più utenti contemporaneamente, agevolandone la lettura.

### **1.3.3 Organizzazione dati e ricerca**

Un altro aspetto interessante di una cartella clinica elettronica è quello di poter organizzare i dati, sia a livello *"orizzontale"*, ovvero a livello del singolo paziente, che a livello *"verticale"*, ovvero raggruppando dati sulla base di uno stesso attributo o caratteristica tra più pazienti che costituiscono una popolazione. Questa funzionalità risulta di particolare utilità per una ricerca

rapida, precisa ed efficiente di dati anche in database molto complessi, dalla cui analisi è possibile ottenere informazioni di utilità, sia per il singolo paziente (malattie pregresse o farmaci assunti...), che per l'intera popolazione, in quanto permette l'osservazione di dati quali trend ed evoluzioni epidemiologiche utili ai fini di ricerca e di sanità pubblica.

### 1.3.4 Caratteristiche di attività

Oltre ai vantaggi precedentemente descritti, il reale potenziale della cartella clinica elettronica risiede nella funzione "attiva" che può assumere. Grazie alla programmazione di algoritmi questo strumento non si limita alla sola registrazione e visualizzazione dei dati ma è in grado di intervenire ed interagire in modo proattivo con l'utente o il flusso di lavoro clinico, al fine di generare azioni automatiche o di supporto decisionale.

Alcune delle funzioni attive della cartella clinica elettronica sono le seguenti:

- **Visione integrata dei dati dei pazienti:** permette l'accorpamento di informazioni provenienti da più fonti e sistemi e la loro elaborazione in altre forme, quali grafici ed immagini più semplici da analizzare, in modo tale da avere costantemente una visione complessiva sullo stato del paziente. È anche possibile programmare notifiche o allarmi automatici per avvisare il medico in situazioni di allerta o che necessitano di attenzione, quali possibili allergie, interazioni tra farmaci in via di somministrazione, esiti di laboratorio dai valori critici o più semplicemente un sistema automatico di verifica della copertura vaccinale, di notifica di follow-up programmati e di visite di screening raccomandate.

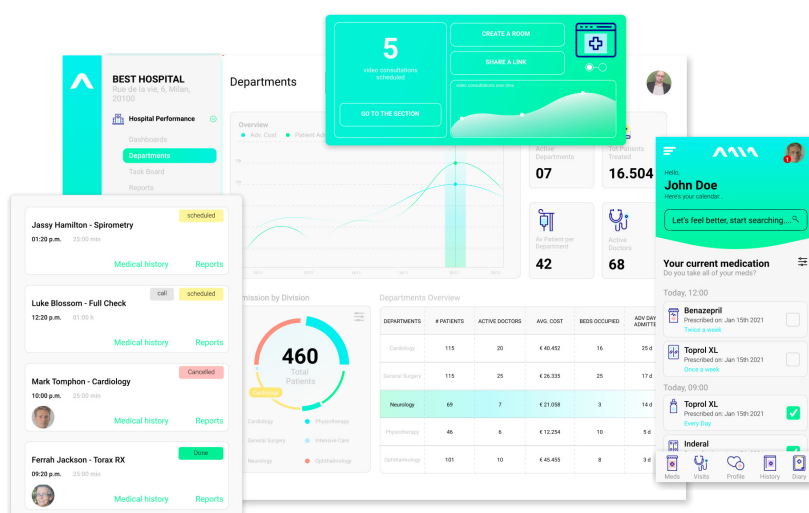


Figura 1.1: Esempio concettuale di interfaccia per la visione integrata dei dati del paziente.

- **Supporto decisionale:** gli algoritmi previsti nei software possono elaborare i dati ricevuti in input nella cartella per suggerire diagnosi differenziali, potenziali protocolli terapeutici

e dosaggi dei farmaci sulla base dei parametri del paziente. In questo modo il medico può essere supportato nella decisione grazie al confronto di diversi quadri e situazioni suggerite.

- **Supporto nell'inserimento dei dati:** attraverso menù a tendina e campi pre-compilati può essere di notevole aiuto a medici ed infermieri nell'inserimento dei dati del paziente, con conseguente risparmio di tempo per la compilazione della cartella, soprattutto nella parte burocratica, aumentando il tempo a disposizione nei confronti del paziente.
- **Accesso a fonti di conoscenza:** la cartella clinica elettronica può agevolare l'accesso a protocolli e bibliografia rilevante riguardante il percorso clinico del paziente, mostrando al medico le più recenti tecniche di cura per il determinato caso o le soluzioni più raccomandate e consolidate.
- **Supporto integrato alla comunicazione:** si possono prevedere dei canali di comunicazione preferenziali tra medici e strutture sanitarie coinvolti nel processo di cura del paziente, al fine di facilitare la comunicazione tra le parti coinvolte. Questi canali possono essere utili anche per comunicare con gli organi amministrativi e burocratici al fine di velocizzare pratiche come quelle per la richiesta di prestazioni specialistiche, di materiali di rifornimento e di rimborsi. Infine, tali canali di comunicazione possono essere usati anche per la comunicazione tra il paziente e le strutture sanitarie, ad esempio agevolando la prenotazione di visite, o tra il medico ed il paziente, programmando notifiche personalizzate sul dispositivo del paziente come quelle per l'assunzione delle terapie prescritte.

## 1.4 Architettura dei sistemi delle cartelle cliniche elettroniche

### 1.4.1 Struttura mainframe

Questo tipo di organizzazione viene anche detta "*general proupose*": consiste nell'utilizzo di un unico software multi-scopo con accesso multi-utente per la creazione e la gestione della cartella clinica elettronica. Questa particolare struttura risulta essere poco vantaggiosa poichè richiede un sistema e una macchina particolarmente complessa, rigida e molo costosa. Risulta anche essere particolarmente vulnerabile in quanto superate le barriere di sicurezza della singola macchina si ha accesso completo a tutte le informazioni di tutti i pazienti che vi sono memorizzati. Inoltre, ad oggi, la progettazione di una cartella clinica "*general proupose*" risulta essere quasi impossibile, perchè poco pratico progettare un sistema che sia adatto e compatibile con tutte le differenti esigenze, le applicazioni e gli scopi richiesti dai diversi attori e reparti del settore medico.

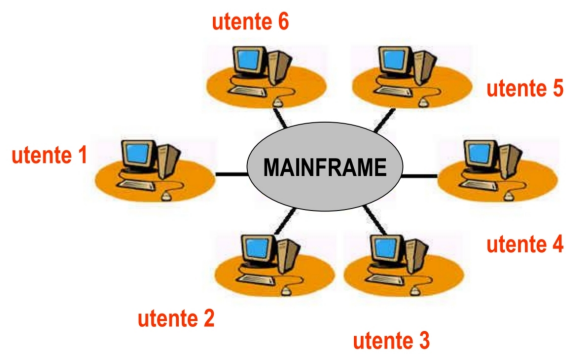


Figura 1.2: Schema grafico della struttura mainframe.[1]

## 1.4.2 Struttura client-server

La maggior parte dei sistemi informativi sanitari sviluppati utilizzano una architettura che predilige una soluzione distribuita e differenziata al fine di poter usare software differenti per ambulatori differenti i quali hanno esigenze e richieste proprie.

In questo tipo di struttura sono presenti diverse applicazioni specifiche per ogni ambito di interesse ed un sistema centrale che le mette in comunicazione. In questo modo le procedure e le applicazioni rimangono proprie del reparto e del servizio erogato in quanto solo le informazioni sono condivise attraverso una rete, solitamente in locale (LAN), che registra i dati in un database centrale.

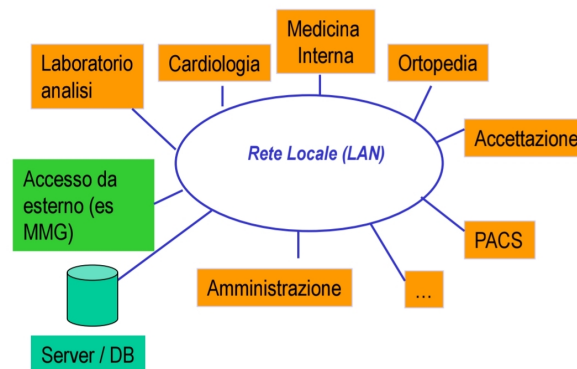


Figura 1.3: Schema grafico della struttura client-server.[1]

Al fine di poter integrare e condividere le informazioni tra le varie componenti del sistema, è necessario utilizzare degli standard di comunicazione che permettano l'interoperabilità dei dati. Gli standard sono delle regole di comunicazione che forniscono un modello comune per la registrazione dei dati. Questi standard possono riguardare la *semantica* e la struttura delle informazioni, la *codifica* della terminologia medica e la *sintassi* di comunicazione. In particolare un sistema si dice *integrato* quando tutti i componenti del sistema utilizzano software che rispetta le stesse regole di comunicazione e di codifica dell'informazione dell'intero sistema, permet-

tendo così uno scambio diretto dei dati; si dice invece *non integrato* quando, non utilizzando uno stesso linguaggio di codifica, sono necessari software detti *middleware* che "traducono" le informazioni nei linguaggi comprensibili alle parti. I vantaggi relativi a questa configurazione sono una maggiore sicurezza per i dati del paziente e una qualità migliore dei dati registrati in quanto è più facile evitare errori di ridondanza, di codifica o di inconsistenza.

## 1.5 Il Fascicolo Sanitario Elettronico

Progettare una cartella clinica elettronica unica per il paziente risulta una soluzione irrealizzabile in quanto non riuscirebbe a soddisfare le necessità di tutte le strutture sanitarie con cui il paziente è destinato ad interagire nell'arco della sua vita. Per questo motivo attualmente risulta vantaggioso lo sviluppo del fascicolo sanitario elettronico.

Il D.L. 179/2012 definisce il Fascicolo Sanitario Elettronico (FSE) come l'insieme dei dati e documenti digitali di tipo sanitario e sociosanitario generati da eventi clinici presenti e trascorsi, riguardanti l'assistito. Secondo la normativa, il FSE è istituito dalle regioni e province autonome ai fini di prevenzione, diagnosi, cura e riabilitazione, di studio e ricerca scientifica in campo medico, biomedico ed epidemiologico, di programmazione sanitaria e di valutazione della qualità dell'assistenza sanitaria. [6] Questa soluzione prevede l'archiviazione dei dati del paziente relativi ad una visita o ad una prestazione direttamente nel database della struttura sanitaria nella quale questi sono raccolti. All'interno della singola struttura viene quindi istituito quello che dal Garante per la Protezione dei Dati viene definito Dossier Sanitario Elettronico. [7]

All'interno del dossier viene documentata la storia clinica del paziente presso quella singola struttura. Le informazioni vengono poi messe a disposizione per la costruzione del fascicolo sanitario elettronico. In questo modo, quella a cui comunemente ci si riferisce come cartella clinica elettronica del paziente diventa il fascicolo sanitario elettronico, ovvero un deposito che contiene gli indirizzi, o puntatori, che conducono a delle cartelle cliniche elettroniche più piccole, conservate localmente dall'azienda sanitaria in cui sono state prodotte. Con questo sistema si ha un "indice" centrale che tiene conto dell'ubicazione di un particolare tipo di informazione del paziente. Senza questo indice, volendo consultare un particolare dato, sarebbe necessario eseguire una query globale su tutte le possibili fonti distribuite sul territorio, rendendo l'approccio non praticabile. Grazie alla presenza di questo indice invece, consultando la cartella clinica di un determinato paziente, nei limiti dei privilegi di accesso alle informazioni, vengono generate una serie di queries che individuano le locazioni di tutte le informazioni relative al paziente. I risultati di queste interrogazioni vengono aggregate tra loro in tempo reale al fine di produrre la cartella clinica completa del paziente.

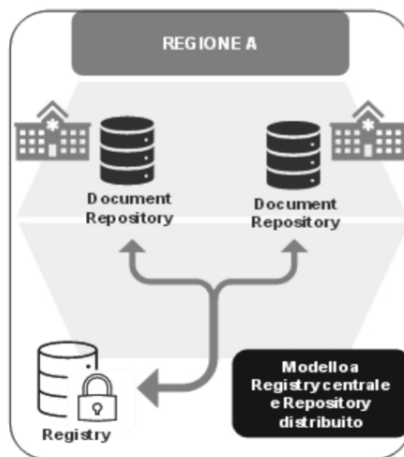


Figura 1.4: Struttura del fascicolo sanitario elettronico a livello regionale.[8]

Per aggiungere una nuova informazione nella cartella clinica, il dato viene memorizzato nel sistema informativo dell'azienda erogatrice del servizio sanitario ed un nuovo link a questa informazione viene aggiunto nell'indice del fascicolo elettronico.

Questo approccio risulta essere favorito dalle aziende sanitarie in quanto possono "esercitare il controllo" sulle informazioni da loro prodotte favorendo l'utilizzo di software differenti per la gestione delle cartelle cliniche "locali", con una programmazione ad hoc, adatta alle esigenze delle strutture e degli ambulatori nella quale vengono utilizzate.

Tuttavia questa soluzione presenta anche alcune criticità. In primo luogo, con l'espandersi delle informazioni contenute nel fascicolo elettronico di un paziente, è necessario un numero sempre maggiore di queries, rendendo l'operazione di assemblamento della cartella meno efficiente perchè maggiormente complessa ed onerosa. Inoltre, tale metodo richiede che ogni sistema informativo locale sia costantemente connesso con quello globale, con conseguente maggiore esposizione e vulnerabilità a possibili intercettazioni ed accessi illeciti alle informazioni. Inoltre, questa metodologia risulta critica per la ricerca di dati per analisi della popolazione, ai fini della sanità pubblica. Per individuare tutti i pazienti che soddisfano una determinata condizione è infatti necessario assemblare la cartella clinica di ogni singolo paziente appartenente alla popolazione in esame con le informazioni provenienti da ogni sorgente e solo successivamente è possibile esaminare singolarmente ogni cartella clinica. Questa operazione di ricerca sequenziale ha un tempo computazionale critico che cresce linearmente con la dimensione della popolazione in esame. Infine, per questa applicazione, risulta ancora più importante l'utilizzo di standard di comunicazione che garantiscano l'interoperabilità tra le cartelle generate da sistemi con software e hardware differenti. Semplificando la comunicazione tra sistemi è anche possibile operare una pre-indicizzazione delle informazioni contenute nel fascicolo, al fine di agevolare le osservazioni dei dati ai fini di ricerca. Lo standard internazionale attualmente maggiormente

diffuso è **HL7**, Health Level 7.



## **Capitolo 2**

# **Possibili anomalie nelle cartelle cliniche elettroniche e rischi connessi**

### **2.1 La problematica**

Le cartelle cliniche elettroniche ricoprono un ruolo di fondamentale importanza. I dati che vi sono contenuti, infatti, offrono promettenti opportunità come base per ricerche e studi in ambito medico e clinico, per l'erogazione del servizio di assistenza sanitaria al paziente e per politiche in ambito sanitario. Le loro potenzialità e la loro utilità sono tuttavia proporzionali alla qualità e all'affidabilità degli stessi dati che vi sono registrati. Infatti, all'interno dei database sanitari, si possono riscontrare delle anomalie costituite da valori che rappresentano qualcosa di incoerente rispetto alla normalità ma che sono una criticità per gli scopi sopra indicati perché rendono i dati inaffidabili e poco sicuri da utilizzare in quanto in grado di alterare i risultati delle ricerche e il percorso clinico del paziente.

Inoltre, l'informatizzazione del sistema sanitario rende le aziende sanitarie sempre maggiormente dipendenti dalle piattaforme digitali e reti di connessione, esponendole alla possibilità di cyber-attacchi e violazioni della privacy dei pazienti. Questi attacchi possono comportare furti di identità, problemi finanziari o gravi conseguenze nell'assistenza sanitaria fornita al paziente. Alcune tipologie di anomalie nelle cartelle cliniche elettroniche possono essere associate ad un intruso che tenta di entrare nel sistema e manipolare le informazioni contenute all'interno dei database clinici.

Infine, le cartelle cliniche elettroniche svolgono un ruolo essenziale per stimare le condizioni di salute di un paziente: la transizione dallo stato di benessere del paziente a quello critico si può osservare nel cambiamento nell'andamento dei parametri vitali nel tempo. Spesso però questi cambiamenti sono minimi e non percettibili, venendo quindi definiti come anomalie. Gli attuali sistemi di monitoraggio utilizzano valori soglia per la generazione di allarmi sui parametri vitali

del paziente. Tuttavia, questo approccio causa un elevato numero di falsi positivi e di anomalie non riconosciute in quanto i valori di soglia sono fissi e prestabiliti senza tenere conto della relazione che intercorre tra i diversi parametri vitali. Inoltre, questi sistemi non sono in grado di utilizzare efficacemente le informazioni dovute alla sequenza temporale degli eventi, perdendo importanti informazioni come potenziali trends che potrebbero indicare un lento declino nella salute del paziente.

Sono diverse le possibili tipologie di anomalie che si possono individuare all'interno dei dati contenuti in una cartella clinica elettronica e ogni anomalia comporta diversi significati.

## 2.2 Anomalie da dati implausibili

L'utilità dei dati raccolti nelle cartelle cliniche elettroniche dipende dalla qualità dei dati stessi. Per valutare la qualità di un dato sono state identificate cinque dimensioni o misure diverse: completezza, correttezza, concordanza, plausibilità e attualità. Queste misure indicano quanto i dati raccolti riescono a descrivere in modo accurato l'attuale realtà.

In particolare, viene definita *plausibilità* una dimensione della qualità dei dati della cartella clinica elettronica che indica quanto questi siano credibili.[9]

Viene quindi valutata la possibilità con la quale un determinato dato riportato nella cartella clinica si possa manifestare concretamente, andando a considerare quanto questo sia coerente con la realtà, con il contesto fornito da altri dati e con l'attuale conoscenza o aspettativa medica.

### 2.2.1 Esempi

Si consideri ad esempio una cartella clinica che riporta un'età anagrafica del paziente pari a 150 anni. Tale dato è implausibile e risulta sicuramente essere un'anomalia. Questa età infatti risulta impossibile da riscontrare nella realtà e sarà certamente incoerente con un'eventuale data di nascita correttamente registrata.

È importante sottolineare che la plausibilità di un dato non implica necessariamente che tale dato descriva la realtà. Ad esempio, l'indice di massa corporea, o BMI, è un coefficiente utilizzato in ambito medico e nutrizionale per stimare se il peso di un paziente è proporzionale alla sua altezza. Tale indice viene calcolato attraverso la formula

$$BMI = \frac{\text{peso [kg]}}{\text{altezza}^2 [m]}$$

Un BMI pari a  $15 \frac{kg}{m^2}$  risulta quindi essere un valore plausibile, tuttavia è anche non veritiero se all'interno della stessa cartella clinica si riscontrano un peso di 70 kg per un'altezza di 170 cm.

L'indice infatti dovrebbe risultare pari a circa  $24 \frac{kg}{m^2}$ .

Allo stesso modo un dato non plausibile non è da considerare a priori come non veritiero. Un paziente ricoverato per grave ipotermia a seguito di un incidente in montagna può registrare una temperatura corporea critica al di sotto dei  $30^{\circ}C$ . Questa temperatura è sicuramente poco probabile e implausibile, tuttavia il dato, considerando la diagnosi e le condizioni, è veritiero.

### 2.2.2 Possibili cause

La presenza di queste anomalie è spesso dovuta ad errori commessi dal medico o dall'infermiere in fase di registrazione dei dati nella cartella clinica elettronica o a problemi del sistema informatico utilizzato. Infatti, quando questa tipologia di anomalia è verificata come errore nei dati, difficilmente è dovuta ad un fattore pertinente allo stato di salute del paziente.

Al giorno d'oggi si stima che i medici utilizzino almeno il 20% della durata dell'incontro con il loro paziente per inserire nel sistema i dati relativi alla visita stessa. Tra i fattori che influenzano la crescente quantità di tempo richiesta per questa operazione ci sono la crescente quantità di dati richiesti nei software che gestiscono le cartelle cliniche, ai fini di sfruttarne tutte le potenzialità, e un'interfaccia software non sempre "user-friendly" che agevoli e velocizzi l'operazione di inserimento dati al medico. È ragionevole quindi pensare che durante questo processo il medico possa commettere errori nell'inserimento dei dati a causa di distrazione, errata digitazione, errata interazione con il software, dialogo con il paziente o altri fattori che possono interferire con questa operazione.

Altri fattori che possono influenzare la presenza di dati non plausibili nella cartella clinica sono la quantità di informazioni limitate o inaccurate in merito al paziente, problemi di sistema durante il processo di registrazione ed elaborazione dei dati o sistemi di codifica e versioni di software utilizzati non aggiornate allo stato dell'arte.

### 2.2.3 Possibili strategie di riconoscimento automatico

Alcune di queste anomalie, soprattutto se riguardante dati in forma numerica, possono essere facilmente rilevate attraverso l'utilizzo di semplici algoritmi che verificano se i dati forniti in ingresso alla cartella clinica elettronica rispettano determinate condizioni.

Alcune di queste condizioni ad esempio sono:

- **Range di appartenenza:** è un tipo di controllo che viene eseguito su valori di tipo numerico. Questo infatti verifica che il valore numerico in questione sia compreso in un determinato intervallo. Trattandosi di un ambiente medico infatti, è ragionevole pensare

che molte misure eseguite siano comprese in un determinato intervallo fisiologicamente ammissibile.

- **Pattern del dato:** è un controllo che verifica se il dato sia stato inserito sotto forma di un'espressione corretta intesa come forma del dato: il dato deve avere un numero massimo o minimo di cifre, deve essere testuale o con determinati caratteri ammissibili.
- **Compatibilità computazionale:** verifica che il dato inserito sia compatibile con la relazione matematica prevista per quel campo specifico o per altri campi che vi sono associati. Per esempio nell'inserimento di un referto dell'analisi del sangue, la composizione totale dei diversi tipi di globuli bianchi presenti nel sangue deve risultare pari a 100.
- **Consistenza:** è un controllo che permette di individuare errori confrontando la compatibilità logica fisica tra più parametri riportati in ingresso; un controllo di questo tipo fornisce un errore, ad esempio, quando individua il report di una diagnosi di tumore alla prostata in un individuo femminile.
- **Coerenza temporale:** è un controllo che verifica che non ci siano drastiche ed improbabili differenze tra i valori della misura appena eseguita e i valori della stessa misura eseguiti precedentemente. Un errore di questo tipo potrebbe essere una differenza di peso di 100 kg in due settimane.
- **Ortografia:** verifica la correttezza ortografica dei dati inseriti in input nella cartella clinica elettronica.

Nel capitolo 3 verranno approfonditi due algoritmi, "FindFPOF" (sez. 3.1) e un algoritmo basato su autoencoder (sez. 3.2), utilizzati per la ricerca di questo tipo di anomalie.

## 2.3 Anomalie da accessi illeciti

Il sistema informativo sanitario digitale non è esente da esposizione a cyber attacchi. Accessi illeciti ai dati del paziente, se trascurati, possono comportare gravi problemi per il paziente e per il medico curante. Un'anomalia del rischio per la sicurezza si verifica quando personale non autorizzato accede ai dati contenuti nelle cartelle cliniche elettroniche attraverso credenziali rubate o quando personale autorizzato utilizza le proprie credenziali per scopi non leciti. La persona che accede a questi dati, sia essa interna o esterna all'ente in questione, può modificare e manipolare i dati che vi sono contenuti. Queste attacchi comportano violazioni di privacy che possono compromettere i risultati ottenuti attraverso la terapia del paziente.

Sono tre le potenziali fonti di minaccia per un sistema informatico.

1. **Dipendente interno:** un dipendente interno all'ospedale o alla struttura sanitaria, dotato quindi delle credenziali di accesso, usa i propri privilegi per scopi illeciti. Si stima che globalmente, questo tipo di minaccia è responsabile del 35% dei casi di violazione di dati nei sistemi informatici, creando danni maggiori rispetto alle violazioni provenienti dall'esterno della struttura.
2. **Violazione esterna:** una figura esterna alla struttura riesce a rubare le credenziali di un dipendente autorizzato o a violare le barriere di sicurezza del sistema, accedendo ai dati che vi sono contenuti.
3. **Negligenza:** un utente autorizzato utilizza il sistema in modo illecito inconsapevolmente o causa di un comportamento negligente.

Questo particolare tipo di anomalia può essere analizzato attraverso l'utilizzo di algoritmi di intelligenza artificiale e machine learning che, studiando i dati forniti in ingresso, possono individuare eventuali anomalie degne di nota. Individuare e correggere queste anomalie risulta di fondamentale importanza per preservare la sicurezza e la privacy dei dati contenuti nelle cartelle cliniche elettroniche dei pazienti.

Nel capitolo 4 verranno approfonditi "Isolation Forest" (sez. 4.1) e "Local Outlier Factor" (sez. 4.2), due algoritmi progettati per la rilevazione di queste anomalie.

### 2.3.1 Esempio

Il più delle volte, le violazioni ai sistemi informatici provenienti dall'esterno, sono fatte al fine di richiedere un riscatto per la restituzione dei dati ai quali si è ottenuto accesso illecito. Questi attacchi però possono compromettere gravemente l'assistenza sanitaria dei paziente.

Il 10 Settembre 2020 l'ospedale Universitario Tedesco di Düsseldorf è stato oggetto di attacco hacker che ha bloccato tutti i sistemi informatici. L'azione ha compromesso l'organizzazione sanitaria, con gravi ripercussioni sul servizio di assistenza dei pazienti: le operazioni eseguite giornalmente nell'ospedale sono state ridotte dalle usuali 120 a 15, in quanto, nonostante parte della strumentazione medica continuasse a funzionare, i risultati clinici non potevano essere inseriti nel database e nelle cartelle cliniche per poter essere elaborati. In questo episodio si verificò quello che è stato definito il primo omicidio causato da un attacco informatico: a causa dell'indisponibilità dei dati contenuti nella cartella clinica, una signora di 78 anni con danni all'aorta, necessitò di essere trasferita in un ospedale a 40 km di distanza per poter essere operata. La signora morì durante il trasporto.[10]

## 2.4 Anomalie nei parametri del paziente

Questo tipo di anomalia può avere due diversi significati. Il primo caso si verifica quando, nei parametri del paziente, sono presenti dei valori definiti *outlier*. Gli outlier sono parametri "nascosti", lontani dalla maggior parte delle osservazioni eseguite. Significa quindi che il paziente presenta una combinazione di parametri fisiologici difficile da individuare da parte del medico durante una comune osservazione, ma che potrebbe indicare una situazione di criticità per il paziente. Gli attuali algoritmi di intelligenza artificiale e machine learning offrono un elevato potenziale nell'individuazione di questi parametri, con possibilità di segnalazione al medico curante che ne potrà trarre le giuste conclusioni.

Il secondo caso invece, avviene quando l'algoritmo rileva dei dati che indicano una gestione scorretta del percorso di cura del paziente rispetto ai protocolli standard e consolidati previsti per la stessa diagnosi. Questo algoritmo è in grado quindi di individuare potenziali procedure mediche e diagnosi scorrette che possono influenzare in maniera negativa la degenza del paziente.

Nel capitolo 5 verrà trattato un algoritmo basato su "Higher-Order Tensor Network" per l'analisi delle anomalie individuabili nelle sequenze temporali dei valori assunti dai parametri fisiologici.

### 2.4.1 Esempio

L'elettrocardiogramma (ECG) è un esame non invasivo che misura l'attività elettrica del cuore di un paziente. In terapia intensiva, i pazienti sono costantemente monitorati attraverso l'ECG, in quanto l'osservazione delle forme d'onda generate è utile per la diagnostica di malattie cardiache o aritmie letali.

In un organismo sano, l'attività del sistema nervoso autonomo è il principale regolatore della frequenza cardiaca. Le modificazioni dell'attività parasimpatica e simpatica del sistema nervoso determinano quindi fluttuazioni fisiologiche nel ritmo cardiaco; gli intervalli tra due battiti consecutivi (intervalli RR) variano costantemente, in risposta al sistema nervoso che regola in tempo reale la frequenza cardiaca, in funzione di complessi stimoli interni ed esterni. L'osservazione della variazione dei battiti cardiaci, ovvero l'individuazione di anomalie nella sequenza temporale della frequenza, può essere quindi considerata come un promettente biomarcatore fisiologico per la valutazione della gravità e dei possibili esiti clinici in pazienti ricoverati per trauma cranico. In particolare, si osserva che una frequenza cardiaca regolare, ovvero con scarsa o nulla variabilità, è segno di squilibri fisiologici, sviluppi critici o indice di danni al sistema nervoso. Anche l'analisi nel dominio della frequenza dell'ECG è utile a tale scopo. Si individua ad esempio il parametro detto "Potenza Totale", che fornisce una misura della varianza globale degli intervalli RR di un ECG. Si osserva che bassi valori di potenza totale, accompagnati da

un rapporto elevato tra alte e basse frequenze nella composizione del segnale, sono associati a maggiore mortalità nel paziente.[11]

## **2.5 Difficoltà attuali nella rivelazione automatica di anomalie**

Rilevare le anomalie all'interno delle cartelle cliniche elettroniche risulta di fondamentale importanza per aumentare l'affidabilità e la sicurezza con cui utilizzare i dati che vi sono contenuti. Le attuali metodologie utilizzate per tale rilevazione affrontano numerose difficoltà in quanto basate su regole deterministiche che ne impongono notevoli limiti sulla tipologia e complessità di anomalie rilevabili. Inoltre, i parametri che vengono scelti in questi metodi di rilevazione corrispondono alle conoscenze del programmatore o dell'utente e sono limitate al proprio campo di lavoro, impedendo una visione d'insieme necessaria in un ambiente biologico governato da una moltitudine di relazioni complesse tra le variabili.

Negli ultimi tempi si osserva un importante aumento dell'impiego della tecnologia di intelligenza artificiale in quanto è in grado di superare queste limitazioni.

Nei prossimi capitoli verranno esposti alcuni algoritmi sviluppati per la rilevazione automatica delle anomalie nelle cartelle cliniche elettroniche, facendo riferimento alle tre tipologie di anomalie precedentemente esposte: nel capitolo 3 verranno esposti due algoritmi per la rilevazione di dati implausibili, nel capitolo 4 due metodologie per la verifica di accessi illeciti nei dati del paziente e, infine, nel capitolo 5 verrà esposta una tecnologia per la rilevazione di anomalie nella sequenza temporale dei parametri del paziente.



## Capitolo 3

# Rilevazione e analisi delle anomalie determinate da dati non plausibili

Le difficoltà nel rilevare dati implausibili all'interno delle cartelle cliniche derivano dal fatto che non sempre sono presenti degli standard applicabili in ambito clinico che definiscono regole deterministiche valide universalmente per distinguere un valore normale da uno anomalo. In questo capitolo verranno analizzati due diversi algoritmi utili alla rilevazione di queste anomalie. In particolare verranno analizzati l'algoritmo "Find FPOF" (sez. 3.1) e l'algoritmo con *autoencoder* (sez. 3.2). Successivamente verrà analizzata una loro applicazione per la rilevazione di anomalie nei registri oncologici (sez. 3.3).

### 3.1 Algoritmi tipo FindFPOF - Frequent Pattern Outlier Factor

Costituisce uno degli algoritmi non supervisionati utilizzati per l'individuazione di dati non plausibili.[12][13] L'algoritmo calcola un parametro detto *Frequent Pattern Outlier Factor*. L'idea alla base di questo metodo è che dati considerabili normali o plausibili hanno una serie di patterns che si presentano con una frequenza elevata, superiore rispetto a quella dei patterns anomali. Si analizzano di seguito gli elementi determinanti di questo algoritmo:

- $I = \{i_1, i_2, \dots, i_m\}$  è l'insieme di tutti gli  $m$  valori che i parametri del database possono assumere.
- $D = \{t_1, t_2, \dots, t_n\}$  è il data base, contenente  $n$  tuple, ciascuna delle quali è un una combinazione di elementi appartenenti ad  $I$ .

- Una tupla  $t \in D$  si dice contiene un sottoinsieme  $X$  di  $I$  se  $X \subseteq t$ , con  $X$  che viene definito come un pattern di elementi appartenenti ad  $I$ .
- Viene definito supporto di  $X$  la percentuale di tuple in  $D$  che contengono  $X$ , ovvero:

$$support(X) = \frac{||\{t \in D \mid X \subseteq t\}||}{||\{t \in D\}||} \quad (3.1)$$

Il problema consiste quindi nel trovare tutti i patterns frequenti nel data base  $D$  per poi calcolare il punteggio FPOF di ogni elemento. Fornito in ingresso dall'utente una soglia chiamata *minisupporto*, è necessario trovare tutti i pattern con supporto maggiore o uguale alla soglia di minisupporto. I pattern frequenti, ovvero quelli che riflettono la maggior parte dei pattern presenti nelle tuple del database, vengono denotati come:

$$FPS(D, minisupport) = \{X \subseteq I \mid support(X) \geq minisupport\}$$

Per ogni tupla  $t$ , il punteggio di FPOF viene calcolato come:

$$FPOF(t) = \frac{\sum_{X \subseteq t, X \in FPS(D, minisupport)} support(X)}{||FPS(D, minisupport)||} \quad (3.2)$$

Il punteggio appena calcolato rappresenta la percentuale di patterns frequenti presenti all'interno della tupla considerata e pertanto può assumere valori compresi tra 0 e 1. Se la tupla  $t$  considerata contiene un numero elevato di pattern frequenti, anche il punteggio FPOF sarà elevato, prossimo ad 1, e con molta probabilità la tupla non è un valore anomalo. Al contrario, a bassi punteggi, prossimi allo 0, corrispondono valori outlier.

Potrebbe risultare interessante verificare il motivo per cui una tupla viene definita outlier. Questo può essere determinato attraverso l'analisi dei pattern non contenuti nella tupla in esame, ovvero i pattern contraddittori:

$$Contradictness(X, t) = (||X|| - ||t \cap X||) \cdot support(X) \quad (3.3)$$

La formula 3.3 evidenzia come i pattern più lunghi riescano a descrivere meglio le anomalie del database rispetto a quelli più corti; inoltre, maggiore è il supporto del pattern  $X$ , maggiore è il grado di contraddittorietà di  $X$  per  $t$  in quanto un valore di supporto elevato suggerisce una grande deviazione.

## 3.2 Algoritmi basati su autoencoder

Un'autoencoder è una rete neurale artificiale utile per la riduzione non lineare della dimensionalità del database che può essere utilizzato anche per il rilevamento di anomalie nei dati.[14][15] Un autoencoder è costituito da due elementi.

- **Encoder:** trasforma l'input originale in una sua rappresentazione compressa, a cardinalità minore, che occupa meno spazio di memoria.
- **Decoder:** a partire dal codice compresso, ricostruisce l'input nella maniera più fedele possibile.

L'addestramento di algoritmi di autoencoding non supervisionati consiste nel minimizzare le perdite dovute alla procedura, ovvero a minimizzare le differenze tra l'input e l'output ricostruito. In particolare, la rilevazione delle anomalie tramite autoencoder e riduzione della dimensionalità, si basa sul presupposto che se determinati dati hanno variabili correlate tra di loro, allora questi possono essere incorporati in un sottospazio a dimensionalità inferiore in cui i dati normali risultano significativamente diversi dai dati considerati anomali.

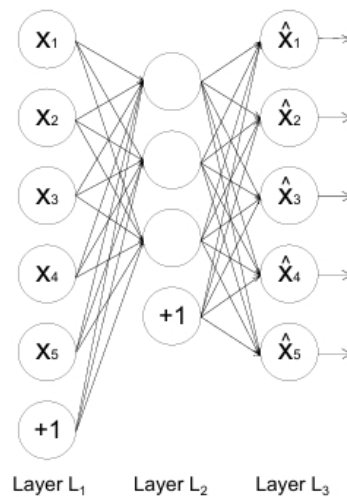


Figura 3.1: Schema grafico di un autoencoder[15]

La figura 3.1 rappresenta sinteticamente il funzionamento di un autoencoder alla quale, nella fase di addestramento, viene fornito all'algoritmo un set di dati pari a  $\{x(1), x(2), \dots, x(m)\}$  in cui ogni elemento è un vettore di  $D$  variabili, cioè  $x(i) \in \mathbf{R}^D$ . Attraverso un numero variabile di neuroni (hidden layer,  $L_2$ ) la dimensionalità dei dati in input viene progressivamente ridotta.

L'attivazione di un neurone  $i$  appartenente ad uno strato  $l$  viene definita come:

$$a_i^{(l)} = f\left(\sum_{j=1}^n W_{ij}^{(l-1)} a_j^{(l-1)} + b_i^{(l)}\right) \quad (3.4)$$

in cui  $W$  rappresenta i pesi tra un neurone ed il successivo, un valore caratteristico della connessione tra due neuroni e che rappresenta quanto forte sia il legame e quindi l'influenza di tale connessione sull'output;  $b$  rappresenta il termine di bias, un fattore proprio del neurone che serve a spostare la funzione di attivazione al fine di modellare la rete neurale e permettere la modellazione di funzioni più flessibili.

In questo algoritmo il punteggio di anomalia per un determinato dato viene rappresentato dall'errore di ricostruzione, calcolato come:

$$Err(i) = \sqrt{\sum_{j=1}^D (x_j(i) - \hat{x}_j(i))^2} \quad (3.5)$$

Tuple con un elevato errore di ricostruzione sono considerate anomalie in quanto sono rare e il punteggio rappresenta elevate differenze nella ricostruzione del valore rispetto a quello atteso.

### 3.3 Una applicazione ai registri oncologici

Gli archivi oncologici contengono informazioni in merito ai pazienti affetti da cancro. Le informazioni che vi sono contenute vengono messe a disposizione dei laboratori di ricerca per monitorare gli sviluppi della malattia nei pazienti al fine di analizzare e migliorare l'efficacia delle terapie utilizzate. Prima di utilizzare tali dati è necessario verificare la loro affidabilità ovvero se sono dati clinicamente plausibili. Le principali cause che rendono un dato non utilizzabile sono l'inadeguatezza o l'inaccuratezza con cui è raccolto e memorizzato per motivi riconducibili al medico o al sistema informatico utilizzato. Lo studio in esame utilizza i due algoritmi precedentemente discussi per rilevare la presenza di dati implausibili nei registri oncologici. Questi algoritmi risultano particolarmente adatti per questo scopo in quanto sono in grado di gestire diverse tipologie di variabili (numeriche, nominali...) senza la necessità di dover conoscere la forma nella quale l'anomalia si presenta.[14]

Lo studio in questione utilizza il *Registro Oncologico tedesco di Rhineland-Palatinate*, con i dati raccolti nel periodo compreso tra gennaio 2019 e ottobre 2021, contenente le cartelle cliniche dei pazienti a cui sono stati diagnosticati tumori al seno, al colon-retto o alla prostata. Ogni

cartella clinica contenie 16 variabili per determinare paziente, diagnosi e procedura prescritta. L'immagine 3.2 mostra un esempio di parte dei dati che possono essere presenti nel registro.

Medical variables		Probability	Plausibility
Sex	Tumor localization		
Male	Prostate	Probable	Plausible
Female	Prostate	Impossible	Implausible
Male	Breast	Improbable	Plausible

Figura 3.2: Esempio di dati di un registro oncologico. [14]

Applicando i due algoritmi discussi ai dati dell'esempio, otteniamo i seguenti risultati.

- **FPOF:** si considerino come patterns frequenti tutti i patterns di lunghezza massima pari a 1; allora ne si possono individuare quattro: {Male}, {Female}, {Prostate} e {Breast} con le rispettive frequenze relative pari a  $\frac{2}{3}$ ,  $\frac{1}{3}$ ,  $\frac{2}{3}$  e  $\frac{1}{3}$ . Calcolando il punteggio FPOF per ogni dato si ottiene:

Sex	Tumor localization	FPOF score
Male	Prostate	1/3
Female	Prostate	1/4
Male	Breast	1/4

Tabella 3.1: Calcolo dei punteggi FPOF dell'esempio nella figura 3.2

Il tumore alla prostata negli uomini è il caso "più normale" e comune e come ci si aspetterebbe l'algoritmo gli assegna un punteggio maggiore; attraverso dei punteggi minori invece, suggerisce gli ultimi due casi come anomalie. Ai fini dello studio, sono stati considerati patterns frequenti quelli con una lunghezza massima di 5 variabili e una frequenza minima del 10% nell'intero set dati.

- **Autoencoder:** se una memorizzazione perfetta dei dati non fosse possibile, l'algoritmo in fase di encoding potrebbe imparare a memorizzare la sola locazione del tumore, ignorando il sesso del paziente; successivamente, in fase di decoding potrebbe ricostruire il sesso del paziente a partire dalla tipologia di tumore. Per un uomo con tumore alla prostata come nel primo caso della figura 3.2, l'algoritmo riuscirebbe facilmente a ricostruire correttamente il sesso, ottenendo un basso errore di ricostruzione e indicando la presenza di un dato plausibile. Al contrario, per il terzo caso, l'algoritmo non riuscirebbe a ricostruire correttamente il sesso del paziente, assegnando il sesso femminile per un tumore al seno, essendo questa configurazione di gran lunga più frequente. L'errore di ricostruzione è quindi elevato e l'algoritmo allerta la presenza di un'anomalia.

In questa applicazione, l'autoencoder presenta tre strati nascosti di neuroni rispettivamente a 16, 8 e 16 nodi.

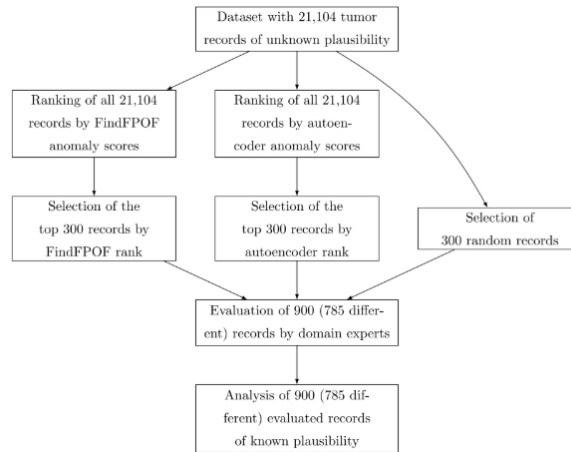


Figura 3.3: Pipeline eseguita nello studio per la validazione di *FPOF* e *Autoencoder*. [14]

L'immagine 3.3 mostra la pipeline che è stata seguita per la validazione dei risultati ottenuti. Lo scopo è di dimostrare l'efficienza con cui gli algoritmi possono individuare le anomalie rispetto alla ricerca manuale eseguita da un team di esperti.

Essendo il data base molto ampio, la sola porzione considerata per il test contiene oltre 21 mila cartelle cliniche, risulta impossibile la sua totale analisi e validazione manuale. Normalmente infatti, la validazione del data set viene eseguita dal team su una porzione di  $k$  valori estratti randomicamente dal database. Lo studio quindi confronta le anomalie rilevate da un campione di questo tipo rispetto a quelle rilevate in un campione determinato dalle prime  $k$  cartelle a cui sono stati assegnati i punteggi di anomalia maggiori da parte dei due algoritmi.

I risultati mostrano come gli algoritmi di rilevazione automatica delle anomalie possano ridurre di un fattore di 3.5 lo sforzo necessario rispetto alla validazione di un campione randomico. Infatti, la percentuale di valori implausibili che vengono rilevati in un campione randomico, cioè la precisione, è pari all'8% ma aumenta al 28% quando la validazione viene eseguita su un campione determinato dalle prime 300 cartelle a punteggi di anomalia più alti. I risultati dimostrano che tale precisione aumenta se riduciamo la dimensione  $k$  del campione considerato. Con un  $k$  pari a 80 infatti il numero di anomalie verificate costituisce il 35% per FindFPOF e il 26% per Autoencoder.

	Records		Tumor localization			
			All	Breast	Colorectal	Prostate
Full dataset	All	$n \left( \frac{n}{n_{tot}} \right)$	21,104 (100%)	<b>11,573 (54%)</b>	6995 (34%)	2536 (12%)
Random sample	Selected	$n \left( \frac{n}{n_{tot}} \right)$	300 (100%)	<b>172 (58%)</b>	87 (28%)	41 (14%)
	Implausible	$\#impl \left( \text{precision: } \frac{\#impl}{n} \right)$	23 (8%)	4 (2%)	<b>16 (18%)</b>	3 (8%)
Autoencoder sample	Selected	$n \left( \frac{n}{n_{tot}} \right)$	300 (100%)	85 (28%)	<b>193 (64%)</b>	22 (8%)
	Implausible	$\#impl \left( \text{precision: } \frac{\#impl}{n} \right)$	83 (28%)	9 (10%)	<b>67 (34%)</b>	7 (32%)
FindFPOF sample	Selected	$n \left( \frac{n}{n_{tot}} \right)$	300 (100%)	40 (14%)	<b>200 (66%)</b>	60 (20%)
	Implausible	$\#impl \left( \text{precision: } \frac{\#impl}{n} \right)$	83 (28%)	3 (8%)	<b>65 (32%)</b>	15 (24%)
All samples	Selected	Total (different)	900 (785)	297 (266)	480 (406)	123 (113)
	Implausible	Total (different)	189 (157)	16 (14)	148 (124)	25 (19)

Figura 3.4: Numero di cartelle selezionate e di anomalie per ogni campione, globalmente e per ogni tipologia di tumore.[14]

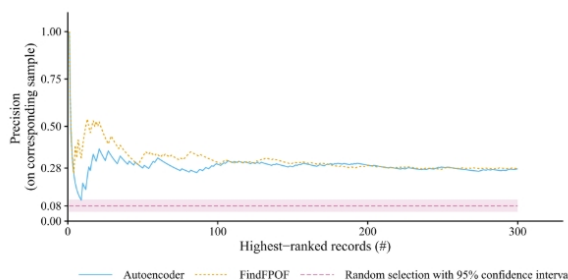


Figura 3.5: Precisione  $\frac{\#impl}{n}$  rispetto al numero di records selezionati a più alto punteggio assegnato.[14]

L'immagine 3.5 mostra che per meno di 100 records selezionati, *FindFPOF* trova un numero maggiore di dati implausibili. Superati i 100 valori la performance di entrambi gli algoritmi è approssimativamente la stessa.

Analizzando i risultati mostrati nella figura 3.6, è stato inoltre osservato che questi algoritmi sono in grado di identificare una varietà maggiore, e sconosciuta, di anomalie rispetto ai metodi di controllo tradizionale.

Variable	Number of distinct values											
	Overall			Breast			Colorectal			Prostate		
	AE	FF	RS	AE	FF	RS	AE	FF	RS	AE	FF	RS
ICD-10 code	21	15	11	6	1	2	13	12	8	2	2	1
TNM T	12	9	10	6	0	2	10	6	6	6	6	3
ICD-O topography	17	13	9	5	1	2	11	11	6	1	1	1
Grading	9	7	7	6	2	2	8	7	6	4	4	2
ICD-O morphology	23	19	6	8	2	2	17	17	4	4	5	1
TNM N	9	5	5	4	0	2	8	4	5	4	4	2
Diagnosis age (binned)	5	5	5	3	3	3	5	5	4	2	4	2
Age at death (binned)	5	5	5	2	2	1	5	5	4	2	4	1
Lateral localization	5	4	4	3	1	3	5	4	4	3	3	1
TNM M	10	7	4	2	1	2	7	5	3	5	5	2
Metastasis	7	7	4	4	3	2	5	5	2	2	4	2
Diagnosis assurance	7	6	3	4	2	2	5	5	2	4	4	1
c/p-prefix T	3	3	3	2	0	2	3	3	3	2	2	2
c/p-prefix N	3	3	3	2	0	2	3	3	3	2	2	2
c/p-prefix M	3	2	3	2	1	1	3	2	3	2	2	2
Sex	2	2	2	1	1	1	2	2	2	1	1	1

Figura 3.6: Numero di valori distinti per ogni variabile in ogni campione di dati implausibili individuati da autoencoder (AE), FindFPOF (FF) e nel campione random (RS).[14]

Le limitazioni di questo studio sono dovute al fatto che gli algoritmi sono stati testati solo su questo set di dati proveniente da un'applicazione concreta. Viste però le dimensioni del test

e l'uniformità nella standardizzazione dei registri oncologici, ci si aspettano risultati analoghi anche in altre applicazioni simili. I prossimi passi per questi algoritmi saranno ulteriori test in ulteriori database, andando a verificare anche l'influenza che i diversi campioni randomici, l'inizializzazione random dei pesi e dei bias della rete neurale e i diversi contesti nell'addestramento della rete neurale possono avere sui risultati dell'esperimento.

## Capitolo 4

# Rilevazione e analisi delle anomalie determinate da accessi illeciti

Gli accessi illeciti possono diventare rilevabili attraverso l'utilizzo di algoritmi di machine learning, in grado di identificare comportamenti anomali nei dati che rappresentano e descrivono gli accessi ai data base e ottenendo risultati accurati ed efficienti. Nel seguente capitolo verranno discussi due degli algoritmi utilizzati per questo scopo. In particolare verrà analizzata la preparazione preliminare dei dati da analizzare (sez. 4.3.1) e verranno considerati l'algoritmo "Isolation Forest" (sez. 4.1) e l'algoritmo "Local Outlier Factor" (sez. 4.2). Verrà infine fatto un confronto tra le performance dei due algoritmi (sez. 4.3).

### 4.1 Isolation Forest

L'algoritmo non supervisionato di Isolation Forest, in breve IForest, costruisce diversi alberi decisionali detti di isolamento che vengono utilizzati per distinguere le anomalie dalle istanze normali.[16][17]

Gli alberi di isolamento prevedono la creazione di alberi binari attraverso la suddivisione successiva del set di dati. Ad ogni ciclo viene scelta casualmente una delle variabili o feature caratterizzanti del data set e un valore di questo parametro, che può essere scelto con casualità uniforme tra i valori assunti dalla variabile o impostato dai ricercatori. Utilizzando tale valore come soglia, si procede con la suddivisione ricorsiva dei dati del database, fermandosi solo quando tutti i dati del nodo appartengono alla stessa classe o quando la lunghezza massima dell'albero è stata raggiunta. In questo modo, attraverso il partizionamento ricorsivo dei dati, si arriverà ad isolare dei punti più o meno velocemente, ovvero si arriverà al nodo foglia dell'albero attraverso un numero più o meno grande di suddivisioni. I punti che vengono isolati più velocemente, ovvero con un numero minore di suddivisioni, sono quelli che con maggiore probabilità costituiscono

un'anomalia. Il concetto alla base di questo algoritmo infatti è che i valori outlier sono valori rari, strani, e pertanto richiedono pochi passaggi per poter essere separati ed isolati da quelli considerabili come normali e che quindi, condividendo molte caratteristiche con la maggioranza, richiedono un elevato numero di partizioni prima di essere isolati. Questa caratteristica rende l'algoritmo più veloce, affidabile e richiede meno memoria, rendendolo particolarmente adatto per database ad alta dimensionalità.

In seguito alla costruzione degli alberi, l'algoritmo calcola per ogni punto il punteggio di anomalia come:

$$s(x, n) = 2^{E(h(n))/c(n)} \quad (4.1)$$

I parametri dell'equazione (4.1) sono:

- $s(x, n)$ : punteggio di anomalia per il punto  $x$  in un database di dimensione  $n$ .
- $h(n)$ : è l'altezza dell'albero decisionale a cui appartiene il punto  $x$ , ovvero il numero di partizioni necessarie per isolare  $x$ .
- $E(h(n))$ : è la lunghezza media attesa di tutti gli alberi decisionali della "foresta" ottenuta dal set di dati.
- $c(n)$ : è una costante che rappresenta la lunghezza media di un percorso di ricerca binaria nell'albero che non fornisce risultato. Viene calcolata come:

$$c(n) = 2H(n-1) - \frac{4(n-1)}{n} \quad (4.2)$$

dove  $H$  è il numero armonico, cioè la somma dei reciproci di tutti i numeri positivi fino ad  $n$ :

$$H(n) = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots + \frac{1}{n} \quad (4.3)$$

---

```

Inputs:  $i$  instance,  $iT$  isolation tree,  $c$  current length
Output: Path Length of instance  $i$ 
1: If ( $iT$  is an external node) then
2: Return  $c + cost(iT.size)$ 
3: End if
4:  $a \leftarrow iT.Normal$ 
5:  $b \leftarrow iT.intercept$ 
6: If  $(i - b).a \leq 0$  then
7: Return Path Length( $i, iT.left, c + 1$ )
8: Return Path Length( $i, iT.right, c + 1$ )
9: End if

```

---

Figura 4.1: Algoritmo per il calcolo della lunghezza del percorso per l'istanza  $i$ . [16]

L'algoritmo IForest assegna quindi ad ogni dato il suo punteggio di anomalia, successivamente normalizzato con la formula (4.5). Il punteggio rappresenta l'unicità del valore rispetto alla maggioranza, pertanto punteggi inferiori indicano la normalità. Solitamente viene considerato come anomalo rispetto alla maggioranza un dato il cui punteggio è superiore a 0.57.

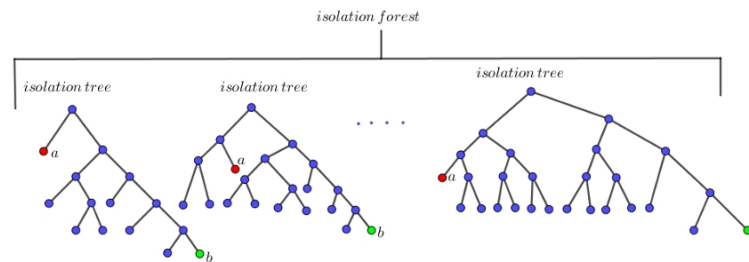


Figura 4.2: Costruzione e rilevazione delle anomalie tramite IForest[17]

La figura 4.2 in esempio, mostra come il valore  $a$ , in qualsiasi albero di isolamento, sia facilmente isolabile dal resto dei dati attraverso percorsi molto brevi. Il punto costituisce quindi un'anomalia. Il punto  $b$  invece, in ogni albero di isolamento, viene isolato attraverso molte decisioni binarie, costituendo quindi un valore considerabile normale.

## 4.2 Local Outlier Factor (LOF)

L'algoritmo non supervisionato di LOF è in grado di individuare le anomalie all'interno dei database confrontando la densità locale calcolata in un punto con la densità dei punti vicini.[17] In questo tipo di algoritmo, un valore anomalo è costituito da un punto che ha bassa densità rispetto ai punti vicini. LOF risulta efficiente nell'individuare valori anomali "locali", ovvero che assomigliano molto ai vicini valori considerabili normali. Nella figura 4.3 si possono osservare le anomalie locali nei punti  $a$ ,  $h$ ,  $i$  e  $j$ ; questi infatti hanno una bassa densità in confronto ai loro vicini e si trovano a breve distanza dai principali cluster  $E$ ,  $F$  e  $G$ . I punti  $b$ ,  $c$  e  $d$  invece sono considerati valori anomali globali in quanto sono a bassa densità rispetto ai punti vicini e si trovano lontani dai principali cluster.

L'algoritmo risulta particolarmente efficace per database ad alta dimensionalità. Considerato un punto appartenente al set dati di input dell'algoritmo, gli ulteriori parametri che l'utente deve fornire in ingresso sono il numero  $k$  di punti vicini da considerare e il numero  $m$  di outlier che si vogliono identificare.

Per ogni punto l'algoritmo calcola la distanza di raggiungibilità dai suoi  $k$  valori più vicini; questa distanza rappresenta la distanza massima tra un punto e i suoi  $k$  punti più prossimi. Successivamente viene calcolata la densità di raggiungibilità locale, ovvero l'inverso della distanza

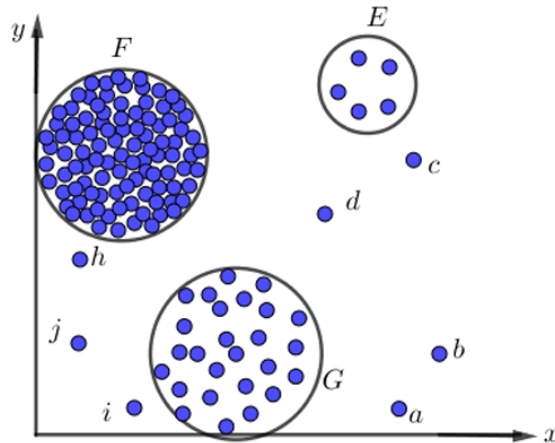


Figura 4.3: Rappresentazione 2D di valori anomali globali e locali.[17]

media di raggiungibilità del punto dai suoi  $k$  punti più vicini. A questo punto viene calcolato il punteggio LOF come:

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} lrd_k(o)}{lrd_k(p) \cdot |N_k(p)|} \quad (4.4)$$

Una volta riordinati in ordine decrescente i punteggi ottenuti, l' algoritmo mostra i primi  $m$  valori dal punteggio più alto, ovvero quelli che implicano la presenza di un possibile valore outlier.

In particolare un punteggio prossimo a 1 indica la normalità, ovvero quando la densità locale raggiungibile è paragonabile a quella media. Al contrario, un valore superiore indica un'anomalia locale in quanto la densità locale è minore della media.

Si osserva che l' algoritmo, andando a considerare sia la densità locale che le distanze tra i vari punti del data set, è in grado di rilevare anche punti anomali che non isolati ma che hanno una densità locale inferiore ai punti vicini.

---

Input:  $k$  - number of nearest neighbors,  $m$  - number of outliers,  $D$  - dataset containing potential outliers  
Output: Top  $m$  outliers  
1: For  $j = 1$  to  $\text{len}(D)$  do  
2: Compute  $k$  nearest neighbor distances ( $k\text{-dist}(p)$ )  
3: Compute neighborhood ( $N_k(p)$ )  
4: End for  
5: Calculate reachability distance ( $\text{reach-dist}_k(p, q)$ ) and local reachability density ( $\text{lrd}(p)$ )  
6: Calculate  $\text{LOF}(p)$   
7: Sort the LOF values of all points in descending order  
8: Return the top  $m$  data objects with the largest LOF values, indicating outliers.

---

Figura 4.4: Algoritmo LOF.[16]

## 4.3 Un'applicazione all'analisi di cartelle in un ospedale inglese

Nello studio analizzato, i due algoritmi vengono testati su un data set costituito dagli accessi alle cartelle cliniche elettroniche di un ospedale dell'Inghilterra Settentrionale.[16]

### 4.3.1 Preparazione dei dati

Prima di poter fornire in ingresso agli algoritmi di intelligenza artificiale i dati da analizzare relativi agli accessi ai data base clinici, è necessario eseguire una fase di pre-elaborazione dei dati.

Nella prima parte di questa fase si puliscono i dati correggendo i valori mancanti ed eseguendo una loro normalizzazione. In particolare, per quanto riguarda i campi vuoti, se la condizione è di contenere un dato *numerico*, allora si provvede alla compilazione con il valore medio assunto da quella variabile nel database; per i dati *nominali* invece, si inserisce la variabile corrispondente alla moda di quel campo.

Successivamente, i dati vengono normalizzati secondo la formula del minimo-massimo:

$$X_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (4.5)$$

Normalizzare i dati affinché i valori siano compresi in un range tra 0 ed 1 permette all'algoritmo un migliore confronto tra i dati, anche tra quelli appartenenti a variabili differenti.

Successivamente viene calcolata la cross-correlazione tra i valori dei diversi parametri secondo la formula:

$$Cross - Correlazione = \sum_n u(n) \cdot h(n - k) \quad (4.6)$$

dove  $u(n)$  rappresenta i valori della prima sequenza di dati nell'istante  $n$  e  $h(n - k)$  rappresenta i valori della seconda sequenza traslati di un valore pari a  $k$ .

Questo passaggio viene eseguito per sfruttare le interrelazioni che intercorrono tra alcune delle diverse variabili presenti nel database, permettendo quindi la riduzione del volume di dati da utilizzare nell'elaborazione della complessità computazionale dell'algoritmo.

La cross-correlazione viene calcolata a coppie tra tutti i parametri stabiliti all'interno del set di dati; le coppie che restituiscono i valori più alti di questo parametro sono quelle che possono essere considerate analoghe e che quindi trasportano informazioni simili. Di conseguenza è

possibile selezionare una sola di queste due variabili e ridurre il volume e la ridondanza dei dati utilizzati.

### 4.3.2 Metriche di valutazione

Per validare l'efficacia degli algoritmi descritti nei paragrafi sopra, sono stati usati diversi parametri:

- **Silhouette score:** è un parametro che viene utilizzato per valutare l'efficienza e la coesione con la quale l'algoritmo divide nei cluster significativi i dati (accesso legittimo o anomalo).

$$Silhouette\ Score = \frac{1}{N} \sum_{i=1}^N \frac{b_i - a_i}{\max(a_i, b_i)} \quad (4.7)$$

In questa formula  $N$  rappresenta il numero totale di punti nel data set,  $a_i$  è la distanza media di un punto dall' $i$ -esimo punto appartenente allo stesso cluster, mentre  $b_i$  è la distanza minima di un punto dal punto  $i$ -esimo appartenente ad un cluster differente. Il range di questo indice è compreso tra -1 e 1. Un punteggio pari a -1 indica un'assegnazione al cluster sbagliato, un punteggio nullo significa sovrapposizione dei cluster mentre un punteggio pari ad 1 significa aver eseguito la giusta assegnazione.

- **Indice di Dunn:** è un'ulteriore metrica di validazione basata sulla distanza minima inter cluster ( $D_{min}$ ) e la distanza massima intra cluster ( $D_{max}$ ):

$$D_{min} = \min\{d(x_i, x_j) | x_i \in C_i, x_j \in C_j, C_i \neq C_j\} \quad (4.8)$$

$$D_{max} = \max\{d(x_i, x_j) | x_i, x_j \in C_k\} \quad (4.9)$$

$$Indice\ di\ Dunn = \frac{D_{min}}{D_{max}} \quad (4.10)$$

### 4.3.3 Risultati

I risultati ottenuti sul database analizzato mostrano una performance migliore dell'algoritmo IForest. Di seguito i valori dei parametri ottenuti.

	Anomalie rilevate	Silhouette score	Indice di Dunn
Isolation Forest	397	0.63	0.45
LOF	358	0.41	0.38

Tabella 4.1: Risultati

In particolare IForest mostra alcune proprietà che lo rendono più performante rispetto ad altri algoritmi per la classificazione delle anomalie. Una tra queste è la scalabilità che lo rende particolarmente adatto per database di grandi dimensioni e ad alta dimensionalità. La chiave di questo algoritmo infatti è la costruzione degli alberi di isolamento. Il numero di nodi interni di questo tipo di alberi è pari a  $(n - 1)$  ed è determinato dal numero di partizionamenti necessari per isolare tutti i punti di un data set. Di conseguenza, il numero totale di nodi è pari alla somma tra i nodi interni e quelli esterni, ovvero  $(2n - 1)$ . Ne risulta che la quantità di memoria necessaria all'algoritmo cresce linearmente con il numero di elementi del data set. L'algoritmo mostra inoltre efficienza computazionale, robustezza e flessibilità sul tipo di dati che riesce ad analizzare. Tuttavia presenta anche alcuni svantaggi, quali il problema di overfitting: si osserva che quando deve lavorare con piccoli data set, con un numero troppo elevato di alberi o con un data set molto clusterizzato, l'algoritmo perde in accuratezza e performance. Inoltre, l'algoritmo richiede l'impostazione da parte dell'utente di diversi parametri quali il numero di alberi da creare e la lunghezza massima che possono avere un impatto significativo sulla performance. Infine, lavorando con parametri casuali in fase di partizione, l'algoritmo soffre di fenomeni di swamping (quando i punti normali sono troppo vicini alle anomalie) e mascheramento (quando nel database è presente un numero elevato di anomalie).

LOF, invece, utilizzando la densità locale dei punti, è particolarmente preciso nell'individuazione di anomalie locali difficilmente rilevabili da altri algoritmi. Risulta particolarmente efficiente anche in data set grandi e ad alta variabilità, anche in presenza di un elevato grado di rumore, ed anche questo algoritmo è in grado di gestire variabili di diverso tipo, sia numeriche che categoriche.

Le limitazioni a questo algoritmo sono la sensibilità alla scelta dei parametri iniziali, una performance che decresce con l'aumentare della dimensionalità e una difficoltà a gestire i data set in cui le anomalie presentano una distribuzione complessa o distribuita tra più cluster. Infine, il fatto di dover calcolare le distanze tra un punto e i suoi  $k$ -esimi vicini lo rende particolarmente oneroso a livello computazionale.



## Capitolo 5

# Rilevazione e analisi delle anomalie dovute a parametri critici o errori clinici

L'utilizzo di algoritmi di intelligenza artificiale consente di creare metodologie di supervisione in tempo reale per la rivelazione di anomalie nell'andamento dei parametri vitali del paziente, in modo da individuare tempestivamente possibili situazioni di criticità da segnalare al medico curante. Nel presente capitolo si analizzerà l'algoritmo di intelligenza artificiale ad addestramento non supervisionato *Higher-Order Tensor Network (AD-HOTN)* (sez. 5.1). Verranno infine analizzati i risultati ottenuti dallo studio in esame (sez. 5.1.3).

### 5.1 Rilevazione delle anomalie tramite Higher-Order Tensor Network

L'algoritmo di rilevazione automatica delle anomalie tramite Rete Tensoriale Multidimensionale utilizza un network per la rappresentante delle interazioni tra i dati e identifica le anomalie monitorando i cambiamenti che avvengono sul grafo quando il dato in esame viene rimosso.[18][19][20]

Di seguito vengono definiti gli elementi fondamentali per l'algoritmo.

- **Sequenza:** è una serie di *eventi*; gli eventi sono rappresentati come dei simboli e sono la più piccola unità dotata di significato di una sequenza. Verranno considerate solo sequenze finite e discrete.
- **Sottosequenza:** è una sequenza ricavata da una sequenza più grande eliminando alcuni eventi ma senza cambiarne l'ordine. Verranno considerate solo sottosequenze contigue.

- **Transizione:** è la più piccola sottosequenza ricavabile da una sequenza, definita come il passaggio tra due elementi contigui della sequenza stessa.
- **Tensore:** è un array multidimensionale ed è una generalizzazione ad ordini superiori di un vettore e di una matrice. In questo studio, per una rete di  $N$  nodi, una transizione viene rappresentata da un tensore a quattro dimensioni:  $\mathcal{T} \in \mathbf{R}^{N \times N \times P \times D}$  (sorgente, target, probabilità di transizione e durata di transizione). Si forma quindi una rete tensoriale denotata con  $\mathcal{N}$ .

I prossimi paragrafi illustrano i passaggi dell'algoritmo. Si inizierà con la rappresentazione delle sequenze temporali di eventi nella rete tensoriale multidimensionale (sez. 5.1.1) per procedere con il calcolo del punteggio di anomalia (sez. 5.1.2).

### 5.1.1 Rappresentazione delle sequenze di eventi nella rete tensoriale multidimensionale

Questo tipo di rappresentazione viene utilizzato per assicurare un'accurata rappresentazione dei fenomeni nascosti in un sistema complesso.

In generale una rete è un grafo rappresentato da  $G = (V, E, \psi)$  con  $V$  che rappresenta i vertici o nodi del grafo,  $E$  che rappresenta gli archi o le connessioni tra i nodi e  $\psi$  è la funzione di incidenza che assegna a ciascun arco le sue estremità.

In una normale rete tensoriale la convenzione prevede che dato un evento, quello successivo è determinato soltanto da quello corrente. In una rete tensoriale multidimensionale, invece, un nodo mantiene una dipendenza multidimensionale con tutti nodi contigui che ne influenzano l'evento, considerando contesti temporali più estesi. Quindi, un nodo di ordine  $k$  denotato come  $n_t | n_{t+1} \dots n_{t-k-1}$  (in cui la transizione al nodo  $n_t$  è influenzata dalle precedenti  $k$  transizioni) ha una probabilità di transizione proporzionale al conteggio  $C(n_t | n_{t+1} \dots n_{t-k-1} \rightarrow n_{t+1})$  tra i due nodi:

$$P(e_{t+1} = n_{t+1} | e_t = n_t | n_{t+1} \dots n_{t-k-1}) = \frac{C(n_t | n_{t+1} \dots n_{t-k-1} \rightarrow n_{t+1})}{\sum_l C(n_t | n_{t+1} \dots n_{t-k-1} \rightarrow n_l)} \quad (5.1)$$

La durata di transizione tra i due nodi viene calcolata come  $t_{s_{t+1}} - t_{s_t}$  e rappresenta la durata del movimento tra i due nodi appartenenti alla sequenza. La durata media di transizione tra due nodi viene allora definita come:

$$AvgD(e_{t+1} = n_{t+1} | e_t = n_t | n_{t+1} \dots n_{t-k-1}) = \frac{\sum_{s \in S} D_s((n_t | n_{t+1} \dots n_{t-k-1}) \rightarrow n_{t+1})}{C((n_t | n_{t+1} \dots n_{t-k-1}) \rightarrow n_{t+1})} \quad (5.2)$$

La probabilità e il tempo di transizione tra due nodi sono parte del peso della connessione stessa e vengono rappresentati nel tensore.

Un passaggio fondamentale nella creazione della rete tensoriale è verificare se incrementare l'ordine di un nodo produce significativa variazione nelle connessioni. Viene assegnato un peso al significato che il tensore in questione assume, al fine di valutare la convenienza nel portarlo avanti. L'elevazione di grado risulta significativo se rispetta il criterio della divergenza di Kullback-Leibler:

$$\mathcal{D}_{KL}(P_{cur}, P_{ExtNode}) > \frac{Order_{ExtNode}}{\log_2(1 + Support_{extNode})} \quad (5.3)$$

La formula indica che la divergenza, ovvero la differenza, nella "distribuzione" del nodo con incremento di multidimensionalità rispetto all'attuale configurazione, deve essere maggiore di una determinata soglia. Tale soglia prende in considerazione l'ordine  $Order_{ExtNode}$  del nodo quando la sua dimensionalità viene aumentata e il supporto  $Support_{extNode}$  dello stesso, ovvero il numero di connessioni passanti per il nodo.

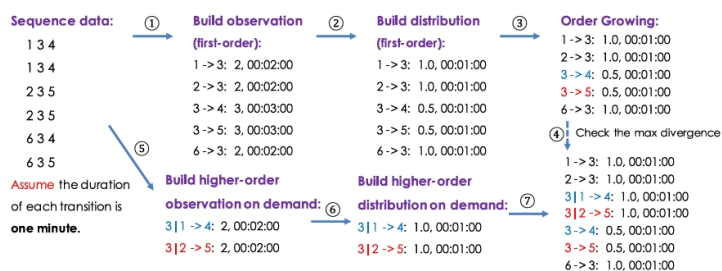


Figura 5.1: Costruzione di una rete tensoriale multidimensionale.[18]

L'immagine 5.1 mostra il processo di costruzione di una rete tensoriale di uno schema dato da sei diversi eventi in sei sequenze diverse; ogni sequenza è formata da tre eventi e la durata di transizione tra ogni evento è assunta pari a un minuto.

- *Step 1:* vengono costruite le osservazioni del primo ordine, contenenti la frequenza (conteggio) di ciascuna possibile transizione e la durata totale di quella transizione.
- *Step 2:* ad ogni connessione viene assegnata la probabilità di transizione e la durata media. Si analizzano due esempi: le possibili transizioni a partire dall'evento 1 sono solo quelle che arrivano all'evento 3, allora la probabilità di questa transizione è pari ad 1 e la durata media, contando due transizioni di questo tipo, è di un minuto. Invece, le transizioni che partono dall'evento 3 sono due, quelle che arrivano all'evento 4 e quelle che arrivano all'evento 5. Allora, la probabilità di ciascuna transizione è pari a 0.5 ed essendoci una transizione per ognuna delle due, la durata media è pari ad un minuto.

- *Step 3*: si calcola il criterio della divergenza di Kullback-Leibler.
- *Step 4*: si osserva che le tre transizioni rimaste in nero in questo step non sono significative a causa del criterio della divergenza; quindi non hanno dipendenze multidimensionali rilevanti e pertanto queste transizioni non vengono sviluppate nei successivi passaggi.
- *Step 5*: vengono esaminate le distribuzioni al secondo ordine dei percorsi scelti.
- *Step 6*: vengono assegnate la probabilità di distribuzione e la durata media dei percorsi validi.
- *Step 7*: non essendoci dipendenze oltre il secondo ordine, il grado massimo è stato raggiunto e quindi la rete tensoriale è stata costruita.

### 5.1.2 Calcolo del punteggio di anomalia

Una volta eseguita la rappresentazione  $\mathcal{N}_s$  del set di sequenze  $S = \{s_1, s_2, \dots, s_n\}$  dei nostri dati, è possibile eseguire l'algoritmo vero e proprio di rilevazione delle anomalie. L'algoritmo consiste nel rimuovere una sequenza  $s_i (i \in (0, n])$  e di costruire la relativa rete tensoriale  $\mathcal{N}_I$ . Iterando il procedimento per le  $n$  sequenze in  $S$ , vengono costruite  $\mathcal{N}_s = \{\mathcal{N}_1, \mathcal{N}_2, \dots, \mathcal{N}_n\}$  reti tensoriali. La sequenza  $s_i$  viene rimossa al fine di osservare i cambiamenti della rete tensoriale rispetto all'originale. Si osserva infatti che, se viene rimossa una sequenza considerata normale, allora la rete tensoriale non subirà grandi modifiche. Al contrario saranno osservabili sostanziali modifiche se ad essere rimossa è una sequenza anomala, in quanto i valori di outliers contengono spesso patterns unici, differenti dai restanti patterns. Eliminare questi patterns significa quindi eliminare alcuni nodi e collegamenti, causando importanti variazioni nella rete tensoriale.

Il cambiamento nella rete viene quantificato attraverso una metrica di distanza che rileva le dissimilarità tra le due reti  $G$  e  $H$ :

$$\mathcal{ND}(G, H) = |E_G \cup E_H|^{-1} \sum_{u, v \in V} \frac{|w_E^G(u, v) - w_E^H(u, v)|}{\max(w_E^G(u, v), w_E^H(u, v))} \quad (5.4)$$

in cui  $w_E^G(u, v)$  rappresenta il peso della connessione tra due nodi  $u$  e  $v$  nella rete  $G$  (rispettivamente per  $H$ ), mentre  $|E_G \cup E_H|$  è il valore di unione degli archi delle due reti. Maggiore è la distanza, maggiore sono le differenze tra le reti tensoriali in questione e pertanto con maggiore probabilità la sequenza rimossa rappresenta un'anomalia.

Per individuare i valori corrispondenti ad un'anomalia è stato scelto di utilizzare un approccio gaussiano in cui la soglia oltre la quale il valore viene considerato anomalo viene calcolata come:

$$threshold = Mean(\mathcal{ND}_S) + 2 \cdot STD(\mathcal{ND}_S) \quad (5.5)$$

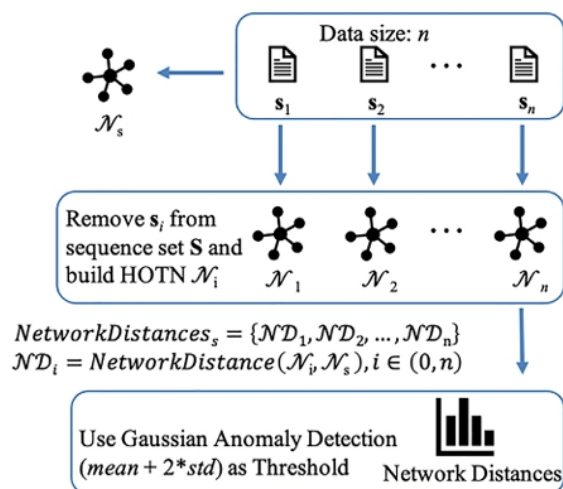


Figura 5.2: Diagramma dell'algoritmo HOTN per la rilevazione di anomalie nelle sequenze temporali.[18]

### 5.1.3 Risultati

Per la validazione dello studio è stato utilizzato un set dati artificiale, creato ad hoc per la sperimentazione. Questo tipo di test offre il vantaggio di conoscere in partenza i risultati desiderati permettendo di valutare con maggiore precisione quelli ottenuti dall'algoritmo.[18]

La metodologia inoltre è stata confrontata con altri algoritmi esistenti per la rilevazione di queste anomalie:

- *kNN*: algoritmo non supervisionato che individua le anomalie attraverso la distanza di un punto dai suoi  $k$  punti più vicini.
- *k-means clustering*: individua le anomalie attraverso un sistema di clustering e di calcolo della distanza di ciascun punto dal cluster più vicino.
- *N-gram*: scompone le sequenze in sottosequenze e verifica la frequenza di patterns rari che vi sono presenti.
- *PST*: utilizza strutture ad albero per il calcolo di suffissi significativi nelle sequenze osservate.

All'interno del data set costruito sono state inserite cinque sequenze anomale. In primo luogo si osserva come l'algoritmo con tensori multidimensionali (AD\_HOTN) riesce ad individuare tutte le anomalie nascoste (figura 5.3).

ID	kNN	k-means	N-gram	PST	ADHON	AD-HOTN
1001	h	m	h	h	h	h
1002	h	h	h	h	h	h
1003	h	h	h	m	h	h
1004	m	m	m	m	m	h
1005	m	m	m	m	m	h

Figura 5.3: Rilevazione delle anomalie nel data set: *h* per identificato, *m* per mancata individuazione.[18]

Successivamente, come parametri di valutazione sono stati considerati precisione, recall e F-score.

Nessuno degli algoritmi utilizzati ha individuato valori falsi positivi di anomalie, pertanto la precisione di tutti gli algoritmi è pari a 1. I risultati dimostrano però una performance superiore per AD\_HOTN in termini di recall, dimostrando valori prossimi al 100% e in termini di F-score. In particolare si ipotizza un F-score superiore per AD\_HOTN preche in grado di rilevare anomalie che contengono durante anormali. Gli altri algoritmi infatti ottengono una performance simile quando le anomalie hanno un tempo di transizione normale.

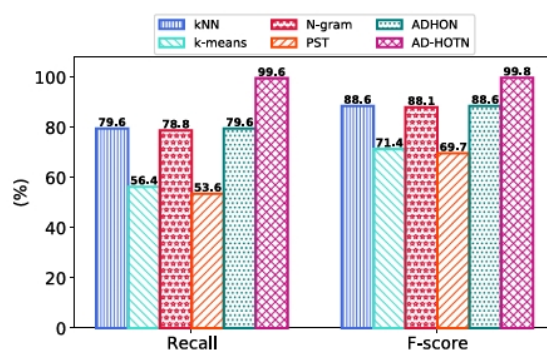


Figura 5.4: Confronto delle performance.[18]

In conclusione si osserva che l'algoritmo è in grado di gestire sequenze di eventi ad alta dimensionalità trasformando le relazioni tra eventi in una rappresentazione grafica multidimensionale. Questo algoritmo può essere utile per identificare ed evidenziare potenziali rischi nella salute di un paziente indicando anomalie nella sequenza di azioni o di parametri del paziente rispetto al trattamento standard.

# Capitolo 6

## Conclusioni

In questo elaborato è stato esplorato il tema della rilevazione automatica delle anomalie nelle cartelle cliniche elettroniche, un ambito di crescente importanza per garantire la sicurezza dei dati, la qualità e l'affidabilità delle informazioni cliniche per la tutela dei pazienti. Attraverso l'esposizione e la discussione di alcuni algoritmi di intelligenza artificiale, si è dimostrato come queste recenti tecnologie possano migliorare significativamente la rilevazione tempestiva di anomalie, riducendo i rischi clinici e le violazioni della sicurezza.

Si evidenzia l'importanza di sviluppare sistemi di monitoraggio intelligenti, capaci di adattarsi alle diverse esigenze ospedaliere e di supportare il personale medico nel processo decisionale.

In prospettiva futura, l'integrazione di tecniche avanzate di machine learning, insieme a un'attenta considerazione degli aspetti etici e normativi non considerati in questo lavoro, potrà ulteriormente potenziare l'efficacia dei sistemi di rilevazione delle anomalie, contribuendo a una gestione più sicura, accurata e affidabile delle cartelle cliniche elettroniche.



# Bibliografia

- [1] G. Sparacino, «Slide e Appunti del corso di Informatica Medica,» Università degli Studi di Padova, A.A. 2024/2025, 2025.
- [2] E. H. Shortliffe e J. J. Cimino, *Biomedical Informatics*. Springer, 2021.
- [3] Unione Europea, *Regolamento (UE) 2025/327 del Parlamento Europeo e del Consiglio sullo spazio europeo dei dati sanitari*, 11 Febbraio 2025. indirizzo: <http://data.europa.eu/eli/reg/2025/327/oj>.
- [4] International Organization for Standardization, *ISO/TR 20514:2005; Health informatics — Electronic health record - Definition, scope and context*, 2005. indirizzo: <https://cdn.standards.iteh.ai/samples/39525/91efcccbdcdf4c4389fc8a6f9ad69f67/ISO-TR-20514-2005.pdf>.
- [5] J. W. Sons, *Electronic Health Record*, 2012. doi: <https://doi.org/10.1002/9781118479612.ch1>.
- [6] Gazzetta Ufficiale della Repubblica Italiana, *Decreto-Legge 18 ottobre 2012, n. 179*, 2012. indirizzo: [https://www.gazzettaufficiale.it/atto/vediMenuHTML?atto.dataPubblicazioneGazzetta=2012-12-18&atto.codiceRedazionale=12A13277&tipoSerie=serie\\_generale&tipoVigenza=originario](https://www.gazzettaufficiale.it/atto/vediMenuHTML?atto.dataPubblicazioneGazzetta=2012-12-18&atto.codiceRedazionale=12A13277&tipoSerie=serie_generale&tipoVigenza=originario).
- [7] Gazzetta Ufficiale della Repubblica Italiana, *Linee guida sul dossier sanitario elettronico (provvedimento n.331)*, 2015. indirizzo: [https://www.gazzettaufficiale.it/atto/serie\\_generale/caricaDettaglioAtto/originario?atto.dataPubblicazioneGazzetta=2015-07-17&atto.codiceRedazionale=15A05443](https://www.gazzettaufficiale.it/atto/serie_generale/caricaDettaglioAtto/originario?atto.dataPubblicazioneGazzetta=2015-07-17&atto.codiceRedazionale=15A05443).
- [8] G. U. della Repubblica Italiana, *Linee Guida per l'Attuazione del Fascicolo Sanitario Elettronico*, 2022. indirizzo: <https://www.gazzettaufficiale.it/eli/id/2022/07/11/22A03961/SG>.
- [9] J. G. K. Hossein Estiri e S. N. Murphy, *A clustering approach for detecting implausible observation values in electronic health records data*, 2019. doi: <https://doi.org/10.1186/s12911-019-0852-6>.

- [10] D. Crossland, «Woman dies after hackers hit AE in Düsseldorf,» *The Times*, 2020.
- [11] P. Zhang, T. Roberts, B. Richards e L. J. Haseler, *Utilizing heart rate variability to predict ICU patient outcome in traumatic injury*, 2020. indirizzo: <https://doi.org/10.1186/s12859-020-03814-w>.
- [12] A. M. Said, D. D. Dominic e B. B. Samir, «Frequent pattern-based outlier detection measurements: A survey,» *2011 International Conference on Research and Innovation in Information Systems*, 2011. doi: 10.1109/ICRIIS.2011.6125705.
- [13] Z. He, X. Xu, J. Huang e S. Deng, «FP-outlier: Frequent pattern based outlier detection,» *Comput. Sci. Inf. Syst.*, 2005. doi: 10.2298/CSIS0501103H.
- [14] P. Röchner e F. Rothlauf, *Unsupervised anomaly detection of implausible electronic health records: a real-world evaluation in cancer registries*, 2023. doi: <https://doi.org/10.1186/s12874-023-01946-0>.
- [15] M. Sakurada e T. Yairi, «Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction,» in *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, Association for Computing Machinery, 2014. doi: <https://doi.org/10.1145/2689746.2689747>.
- [16] M. Tabassum, S. Mahmood, A. Bukhari, B. Alshemaimri, A. Daud4 e F. Khalique, *Anomaly-based threat detection in smart health using machine learning*, 2024. indirizzo: <https://doi.org/10.1186/s12911-024-02760-4>.
- [17] A. M. A. Fadul, *Anomaly Detection based on Isolation Forest and Local Outlier Factor*, 2023. indirizzo: <http://dx.doi.org/10.13140/RG.2.2.17998.43843>.
- [18] H. Niu, O. A. Omitaomu, M. A. Langston et al., «Detecting anomalous sequences in electronic health records using higher-order tensor networks,» *Journal of Biomedical Informatics*, 2022. doi: <https://doi.org/10.1016/j.jbi.2022.104219>.
- [19] M. Saebi, J. Xu, L. M. Kaplan, B. Ribeiro e N. V. Chawla, *Efficient modeling of higher-order dependencies in networks: from algorithm to application for anomaly detection*, 2020. doi: <https://doi.org/10.1140/epjds/s13688-020-00233-y>. arXiv: 1712.09658 [cs.SI].
- [20] J. Xu, T. L. Wickramaratne e N. V. Chawla, «Representing higher-order dependencies in networks,» *Science Advances*, 2016. doi: 10.1126/sciadv.1600028.