Dati Sanitari Sintetici: Compromesso tra Privacy e Utilità nella Ricerca Medica

Autori:
Sara Casadio
Noemi Ferrara
Giorgia Pirelli 000117616

Anno accademico 2025/2026

26 Novembre, 2025

1. Introduzione

Negli ultimi anni, l'Intelligenza Artificiale e il Machine Learning hanno aperto nuove prospettive nella medicina, consentendo diagnosi più precise e terapie personalizzate grazie all'analisi di grandi dataset clinici. Tuttavia, l'utilizzo di dati reali dei pazienti comporta rischi significativi legati alla privacy e alla sicurezza, oltre alla necessità di rispettare normative rigorose come il GDPR europeo e l'HIPAA statunitense. La sensibilità dei dati medici è confermata anche dal loro valore sul mercato illecito: una cartella clinica può arrivare a costare fino a 250 dollari, rendendoli un obiettivo attraente per i cybercriminali.

I dati sintetici si propongono come un'alternativa innovativa. Generati tramite tecniche come le Generative Adversarial Networks (GAN) o i modelli di diffusione, questi dati replicano le caratteristiche statistiche dei dataset reali senza contenere informazioni identificabili sui pazienti. Questo approccio offre la possibilità di sviluppare modelli predittivi efficaci e di condurre analisi approfondite riducendo i rischi etici e legali.

Tuttavia, rimane una questione centrale: è possibile produrre dati sintetici che siano al contempo sicuri e sufficientemente utili? Una protezione della privacy troppo restrittiva può compromettere l'utilità dei dati, mentre dati sintetici molto realistici possono essere vulnerabili a attacchi come il *membership inference*, che rivela se un individuo è presente nel dataset originale, o il *re-identification*, che collega dati sintetici a persone reali.

Diversi lavori precedenti hanno esplorato metodi di generazione e metriche di valutazione dei dati sintetici. Ad esempio, Figueira et al. descrivono vari approcci di sintesi, mentre Hernandez et al. confrontano strumenti di valutazione per determinarne l'efficacia. Il progetto si concentra sul bilanciamento tra privacy e utilità, analizzando quantitativamente come questi due aspetti influenzino le prestazioni dei dati sintetici in contesti sanitari.

1.1. Domande di ricerca

Questo studio si propone di rispondere alle seguenti domande:

- **RQ1:** Quali modelli mantengono le prestazioni più elevate quando addestrati su dati sintetici rispetto alla baseline con dati reali?
- **RQ2:** È possibile conciliare privacy e utilità in un dataset sintetico, o devono essere accettati compromessi significativi?

Per rispondere a queste domande, viene utilizzato l'UCI Diabetes Dataset, un dataset pubblico contenente 768 pazienti con variabili mediche predittive e diagnosi di diabete.

L'analisi si sviluppa attraverso quattro fasi. Innanzitutto, vengono generati dataset sintetici con diversi livelli di protezione della privacy: nessuna privacy, privacy moderata e privacy forte. Successivamente, viene valutata la somiglianza statistica confrontando distribuzioni e matrici di correlazione tra dati reali e sintetici mediante test statistici standard. Nella terza fase, viene verificata l'utilità dei dati sintetici addestrando modelli predittivi che imparano a diagnosticare il diabete valutandoli su dati reali attraverso le metriche Accuracy, Precision, Recall, F1-Score e ROC-AUC. Viene poi implementato un attacco di *membership inference* per testare la resistenza dei dati sintetici e individuare eventuali fughe di informazioni. Infine, il compromesso tra privacy e utilità viene visualizzato attraverso grafici che mostrano come le diverse configurazioni influenzino le prestazioni, permettendo di identificare il punto di equilibrio ottimale.

2. Generezione dati sintetici