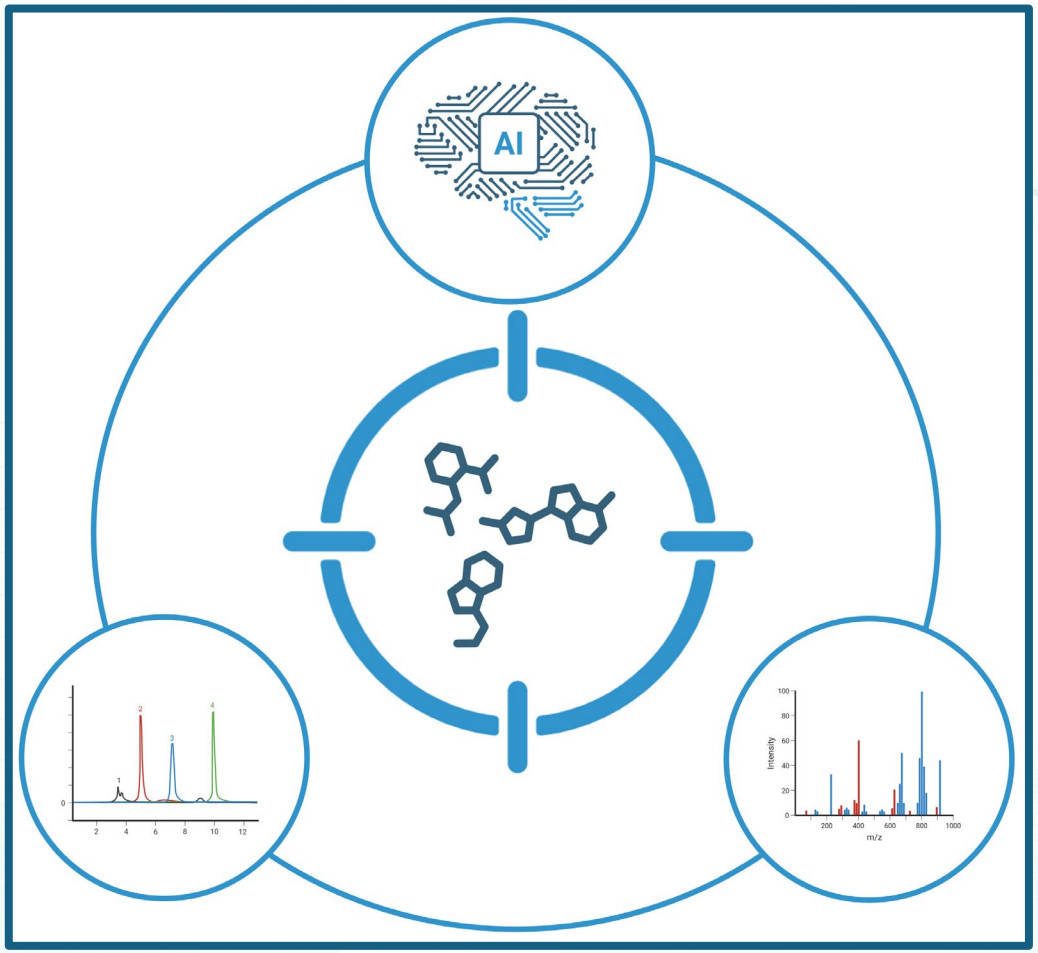
MS UniTE LAB

UNITE UNIVERSITÀ DEGLI STUDI DI TERAMO



Diagnosi 4.0

Quando l'intelligenza artificiale incontra l'analisi molecolare

Federico Fanti, PhD

Responsabile Laboratorio di Spettrometria di Massa
Dipartimento di Bioscenze e Tecnologie Agro-Alimentari ed Ambientali
Università degli Studi di Teramo

Teramo, 10/06/2025

Cosa significa fare una diagnosi?

Fare una diagnosi significa **riconoscere una malattia** a partire da **segnali, sintomi, dati e indizi**, e attribuirle un **nome clinico**, una **causa probabile** e talvolta una **prognosi**.

La diagnosi classica funziona… ma spesso arriva troppo tardi, con troppi pochi dati e troppe incertezze

Perché serve una nuova diagnosi?

- I dati sono aumentati, ma non sempre vengono utilizzati
- Una singola analisi può produrre migliaia di variabili (es. uno spettro di massa)
- Molti di questi segnali sono informativi, ma difficili da interpretare senza strumenti avanzati
- I segnali biologici cambiano prima dei sintomi
- Molecole presenti in sangue, saliva, urina possono indicare alterazioni fisiologiche molto precoci
- La diagnosi classica arriva dopo l'insorgenza clinica
- Una diagnosi “nuova” deve invece essere predittiva
- I medici non possono gestire da soli questa complessità.
- Non per limiti personali, ma per limiti umani: “Nessun occhio può guardare più di 5.000 molecole contemporaneamente.”
- Serve l’aiuto di modelli intelligenti che vedano pattern nascosti

Perché serve una nuova diagnosi?

DIAGNOSI CLASSICA

Sintomi → segni

Oculare, clinica

Reattiva

Limitata dal tempo

DIAGNOSI PREDITTVA

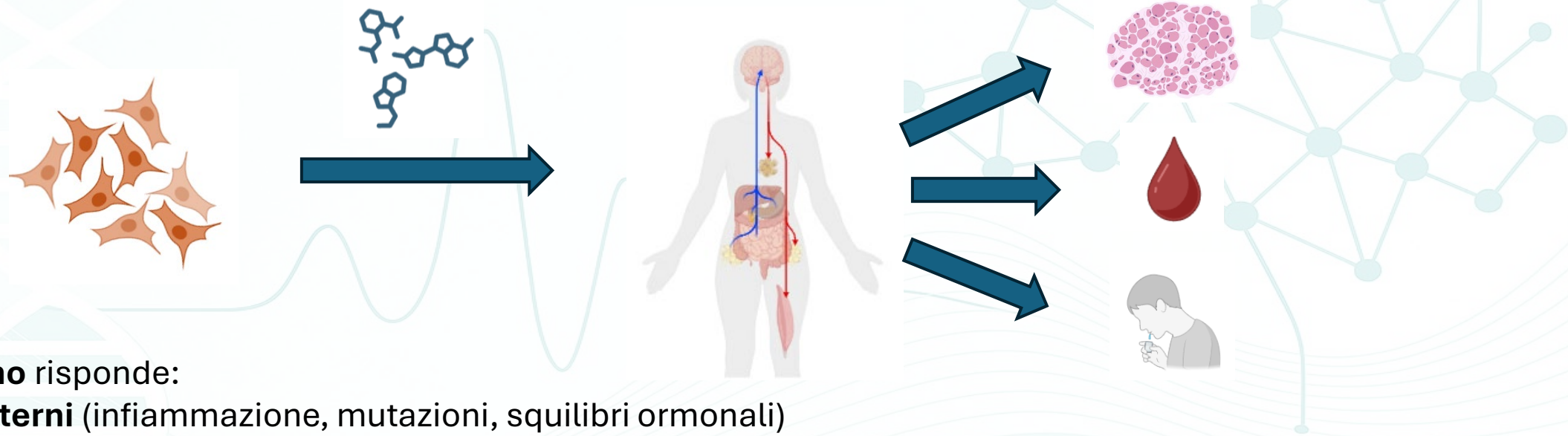
Segnali molecolari → modelli

Molecolare, predittiva

Proattiva

In tempo reale / predittiva

Quando il metabolismo cambia, la malattia inizia a parlare



Il **metabolismo** risponde:

- a stimoli **interni** (infiammazione, mutazioni, squilibri ormonali)
- a fattori **esterni** (dieta, ambiente, farmaci, stress)

Le malattie alterano il **profilo metabolico**

- ogni patologia genera **un'impronta molecolare unica** (o quasi)
- le perturbazioni metaboliche possono **precedere** i sintomi clinici

Come possiamo “vedere” il metabolismo?

Il metabolismo non si vede a occhio nudo → non è un’immagine radiologica, non è un sintomo: è una **rete di reazioni chimiche in movimento**

I suoi segnali sono chimici, piccoli, sfuggenti → metaboliti, prodotti ossidativi, marcatori di stress, SCFA, lipidi, ormoni…

Serve una **tecnologia** che riesca a rilevarli con sensibilità e precisione in microlitri di saliva, gocce di plasma, campioni complessi

Per ascoltare il linguaggio del metabolismo, serve un strumento estremamente sensibile, ad ampio spettro e selettivo:

La spettrometria di massa

Spettrometria di massa: svelare l'invisibile

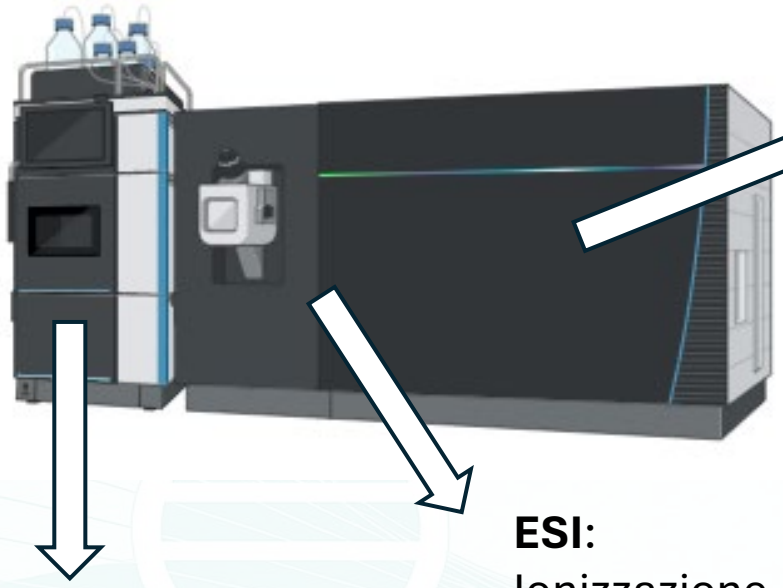
Una tecnica analitica ultra-sensibile che rileva e misura la massa delle molecole in un campione
Serve per "pesare" atomi, molecole, metaboliti con una precisione elevata (es. fino alla 4ª cifra decimale)

Come funziona in 3 passi:

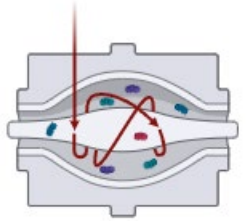
- Ionizzazione → il campione viene caricato (diventa ioni)
- Separazione → gli ioni si muovono in un campo elettromagnetico
- Rilevazione → si misura il rapporto massa/carica (m/z)

Risultato: uno spettro di massa dove ogni picco rappresenta una molecola presente nel campione

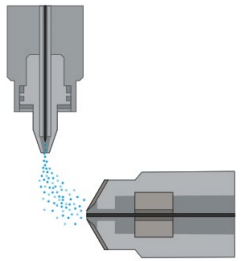
Spettrometria di massa: svelare l'invisibile



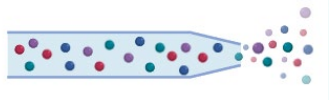
MS:
Separazione in base alla massa

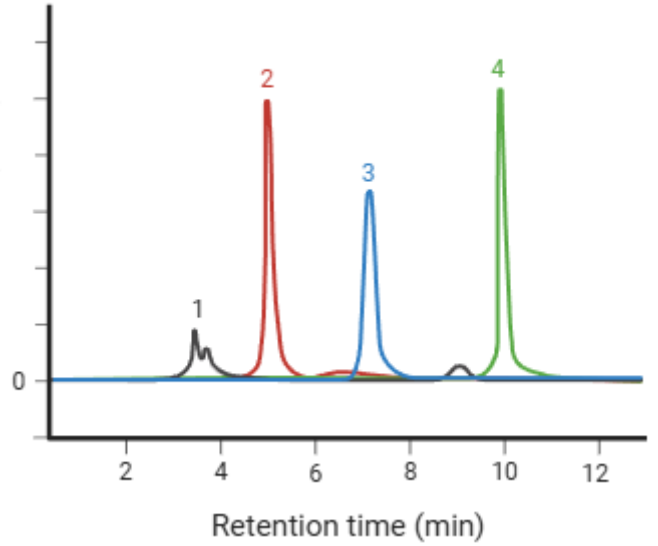


ESI:
Ionizzazione

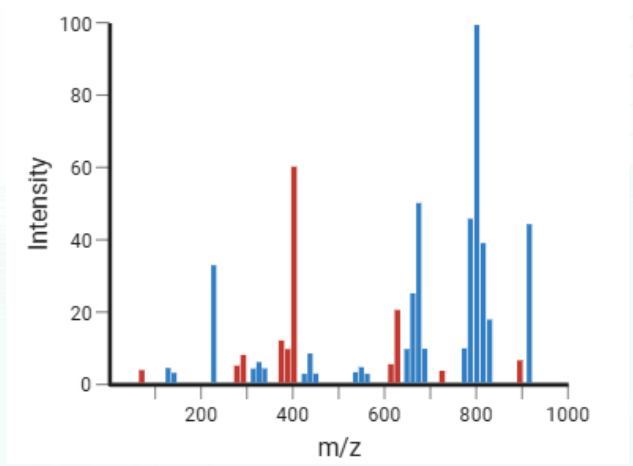


LC:
Separazione in base alla forma





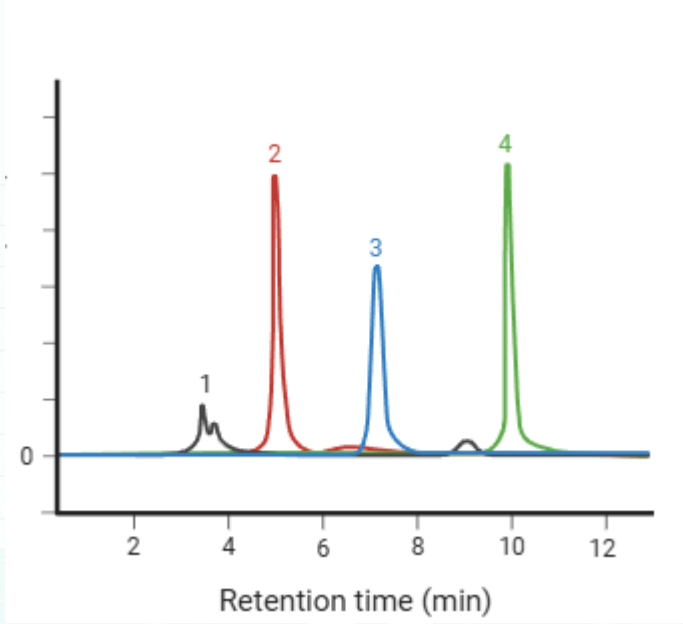
Spettrometria di massa: svelare l'invisibile

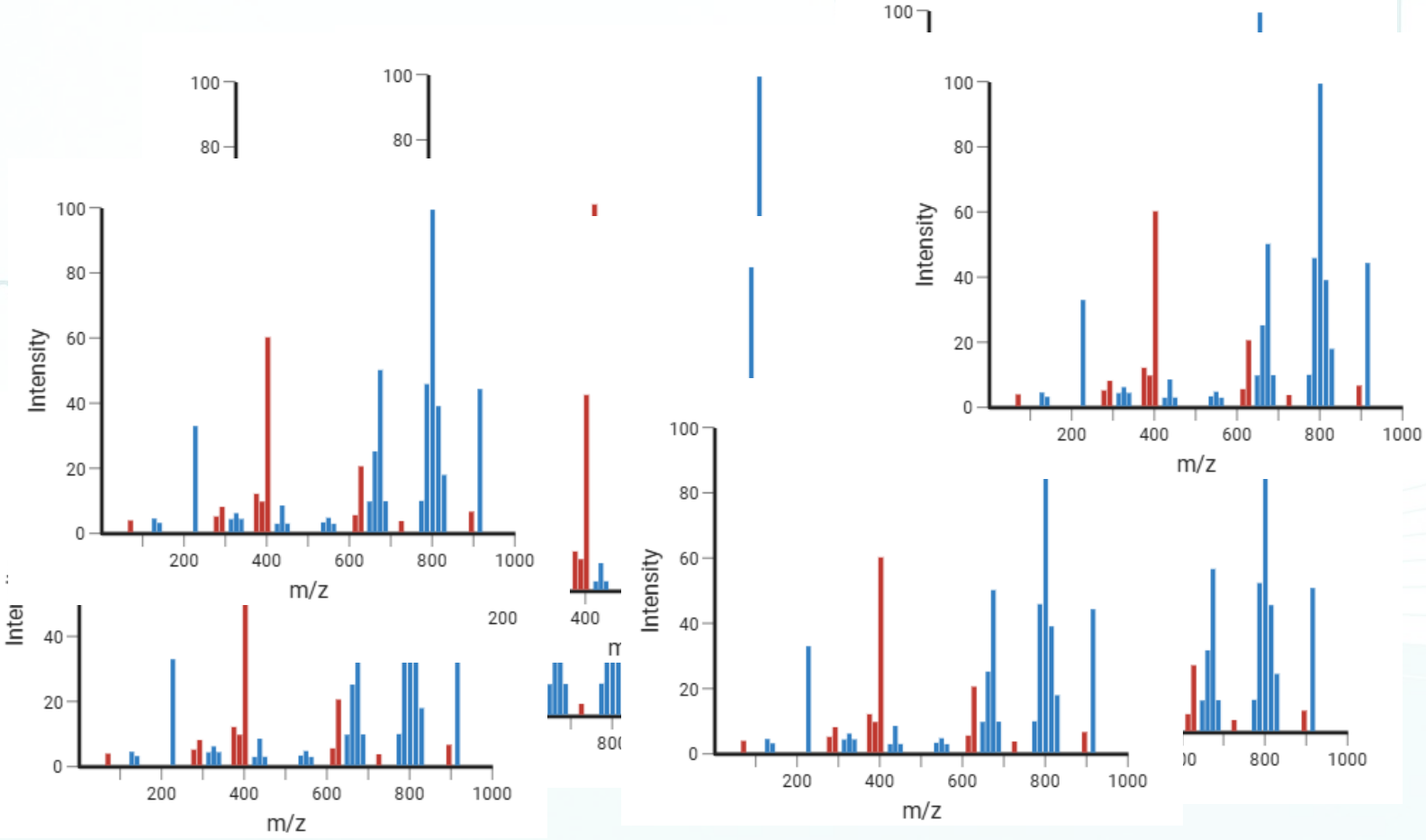


MS UniTE LAB

UNITE UNIVERSITÀ DEGLI STUDI DI TERAMO

Spettrometria di massa: svelare l'invisibile





Ogni campione è una miniera di dati

Una corsa LC-MS produce:

- centinaia di segnali nel tempo (retention time)
- centinaia o migliaia di rapporti m/z
- intensità variabili = concentrazione relativa

Ogni campione è come una matrice quadridimensionale:

tempo x **massa** x **frammenti** x **intensità**

Problema: Come si estrae un significato clinico da tutto questo?

Machine Learning: quando l'algoritmo «impara» dai dati

Il machine learning (ML) è una branca dell'intelligenza artificiale ovvero sviluppa modelli che **imparano automaticamente da dati**, senza regole fisse pre-programmate.

Il modello "comprende" la varianza all'interno dell' analisi e costruisce una funzione interna per riconoscere pattern, classificare, predire

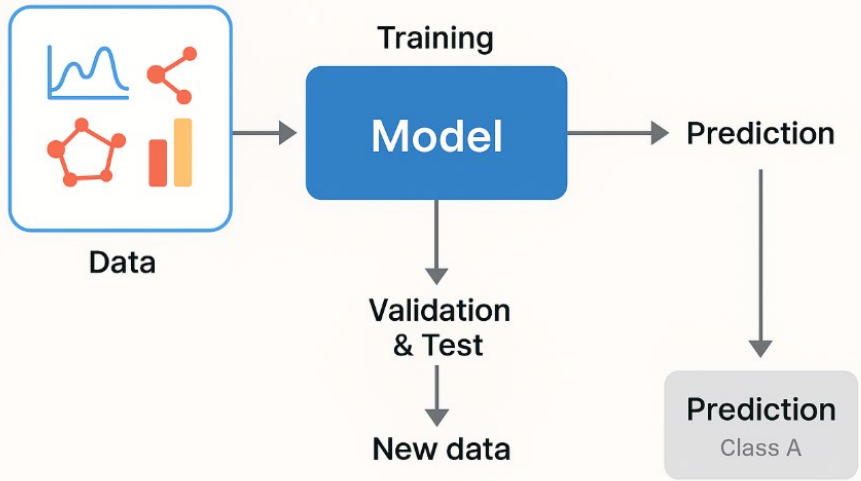
Nell' analisi metabolica:

- Input = spettri, segnali, concentrazioni, profili
- Output = classi (sano vs malato), livelli, rischi, fenotipi

Come funziona il machine learning?

1. Addestramento (training)

- Il modello riceve **esempi già etichettati** (es. campioni con diagnosi nota)
- Impara a riconoscere **pattern ricorrenti**



2. Validazione e test

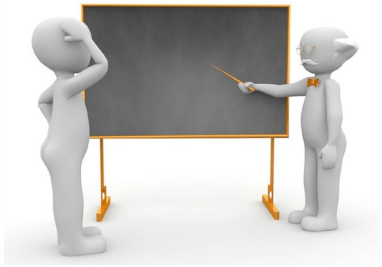
- Si verifica se il modello **generalizza bene** anche su dati nuovi
- Si evitano errori come **overfitting**

3. Predizione

- Il modello analizza **un nuovo campione**
- Restituisce una **classe** (es. sano vs malato) o un **valore numerico** (es. rischio, concentrazione stimata)

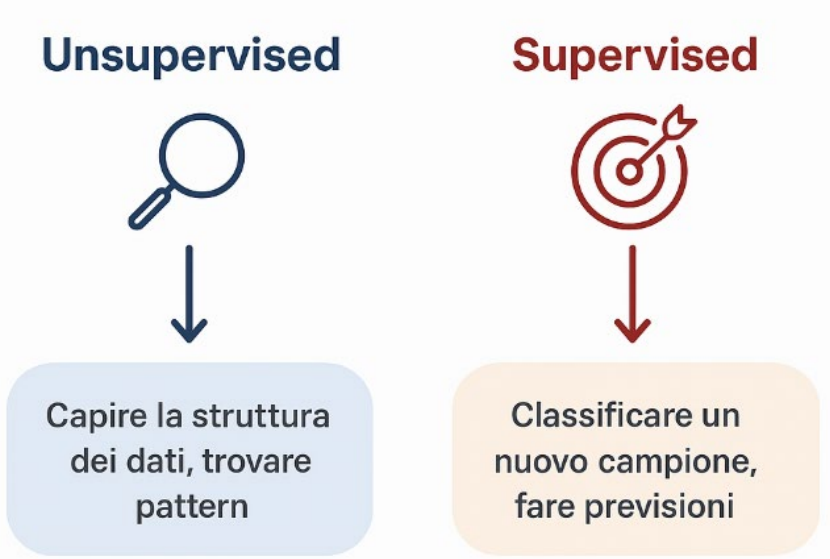
Perché MS+ML è utile?

supervised (vs) unsupervised



Caratteristica	**Supervisionato** (Supervised)	**Non supervisionato** (Unsupervised)
Dati in input	Dati con etichette (classi, valori noti)	Dati **senza etichette**, solo variabili
Obiettivo	Imparare una relazione tra dati e **output noto**	Trovare **pattern nascosti**, gruppi, strutture
Esempio pratico	Diagnosi: "sano" vs "malato" → l'algoritmo impara dai casi noti	Scoperta di sottogruppi metabolici senza sapere chi è chi
Algoritmi tipici	SVM, Random Forest, Logistic Regression, ANN	K-means, PCA, t-SNE, Hierarchical Clustering
Uso in diagnostica	Predizione/classificazione di malattie	Segmentazione di fenotipi, esplorazione dati
Vantaggio	Alta accuratezza se ben addestrato	Può scoprire strutture **inesplorate o inattese**
Svantaggio	Serve una base dati etichettata (clinicamente convalidata)	Risultati meno interpretabili o direttamente utili

supervised and unsupervised



Fase	Tecnica	Obiettivo
🟦 **Esplorazione**	PCA, clustering (unsupervised)	Capire la struttura dei dati, trovare pattern inaspettati
🟧 **Feature selection**	può essere ibrido	Capire quali variabili discriminano tra gruppi
🟩 **Predizione**	SVM, Random Forest, LDA (supervised)	Classificare un nuovo campione, predire un rischio

I modelli non supervisionati: quando i dati parlano da soli

1. Clustering

Raggruppa i campioni simili in base ai loro profili metabolici
Algoritmi: K-means, Hierarchical Clustering, DBSCAN

Esempio: campioni di urina da pazienti (sani e malati) si dividono spontaneamente in 3 gruppi
uno dei tre potrebbe corrispondere a una sottopopolazione clinica non ancora nota.

2. Riduzione della dimensionalità

Semplifica i dati complessi mantenendo le relazioni più importanti
Algoritmi: PCA, t-SNE, UMAP

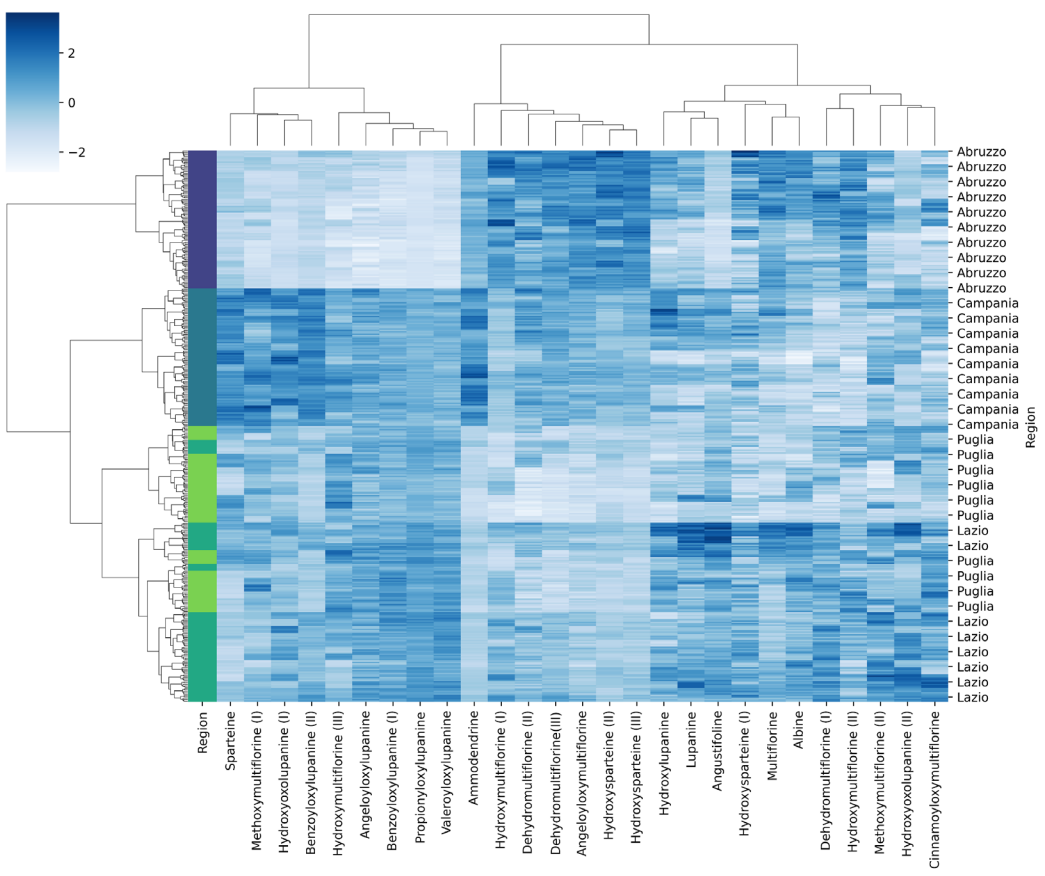
Esempio: un dataset da LC-MS con 1000 variabili viene ridotto a 2 o 3 componenti per essere visualizzato in un grafico semplice
Serve a vedere una separazione tra gruppi.

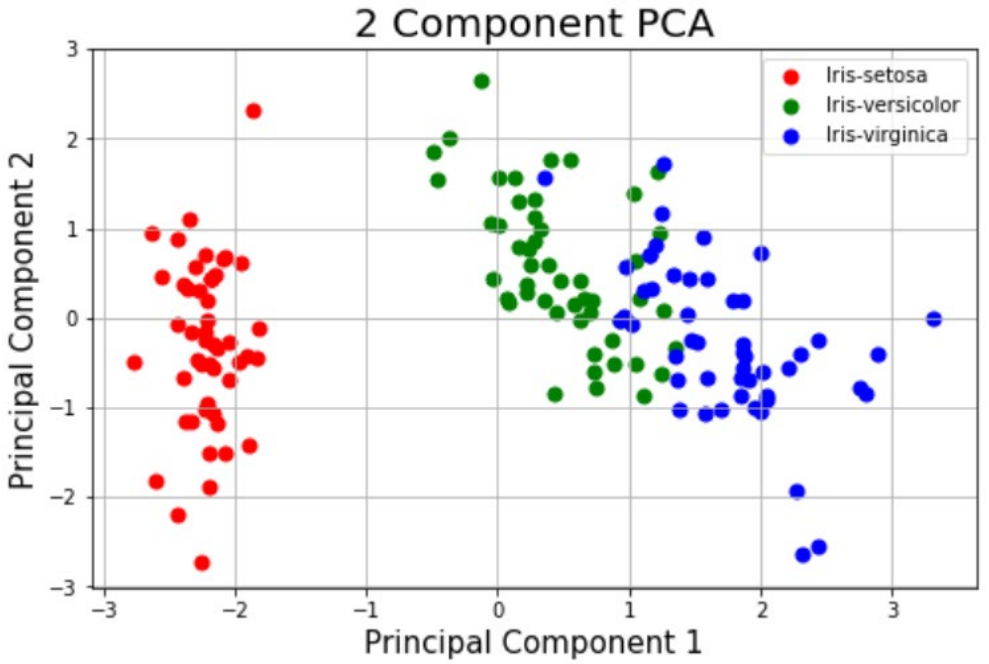
3. Pattern discovery e outlier detection

Scova anomalie o sottogruppi "insospettabili"
Utili per identificare:

- outlier (es. campione contaminato, profilo anomalo)
- segnali deboli ma consistenti

I modelli non supervisionati: quando i dati parlano da soli





I modelli supervisionati: quando sai cosa cercare

1. Classificazione

Prevede una classe o etichetta a partire da un profilo dati

Algoritmi: Random Forest, Support Vector Machines (SVM), Logistic Regression

Esempio: un campione viene classificato come sano o malato

2. Regressione

Prevede un valore continuo

Algoritmi: Linear Regression, PLS, Ridge, Lasso

Esempio: concentrazione di una molecola, punteggio di rischio

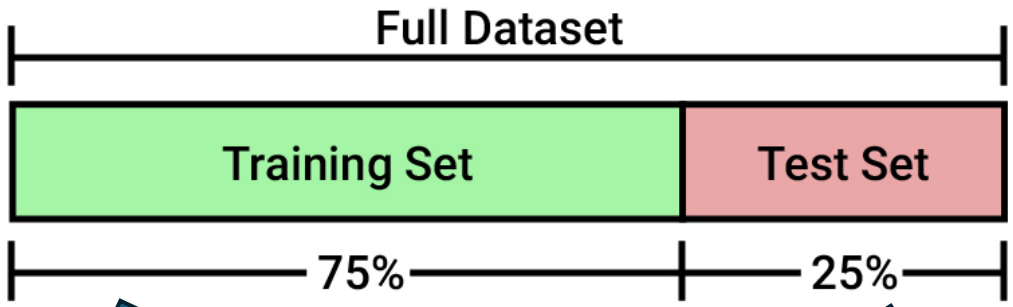
3. Addestramento e generalizzazione

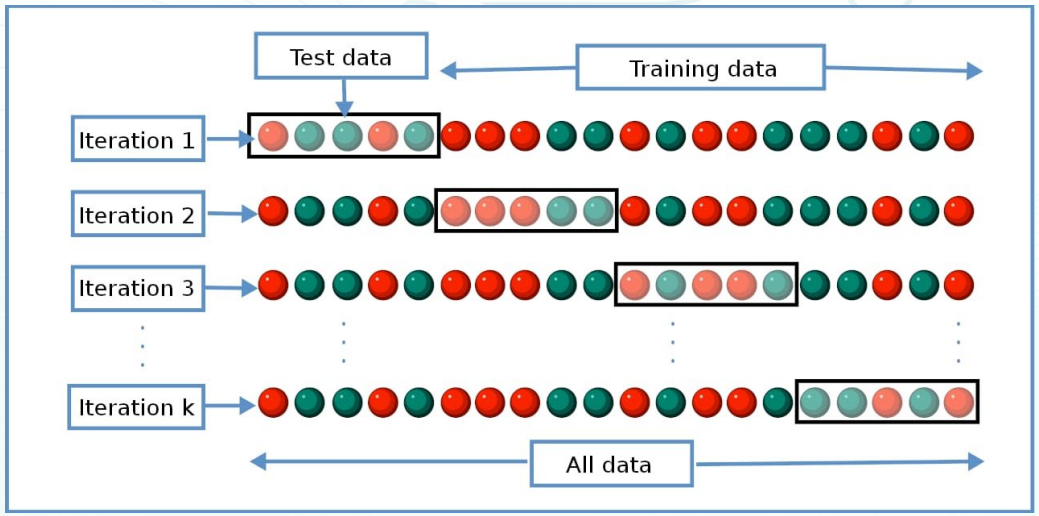
Il modello impara dai dati etichettati

Fa predizioni su nuovi campioni

Se ben costruito, può generalizzare e supportare decisioni cliniche

I modelli supervisionati: Validazione

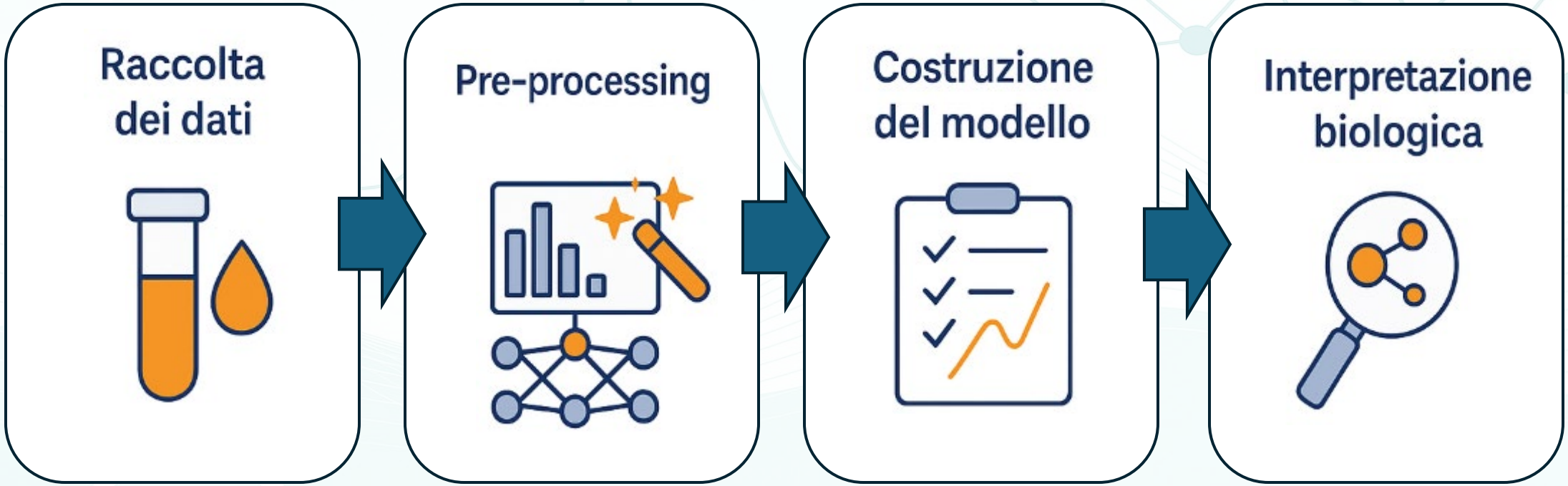




Predicted Class

Actual Class	Positive	Negative	
Positive	True Positive (TP)	False Negative (FN) Type II Error	**Sensitivity** $\frac{TP}{(TP+FN)}$
Negative	False Positive (FP) Type I Error	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Come si applica il machine learning alla diagnosi molecolare?



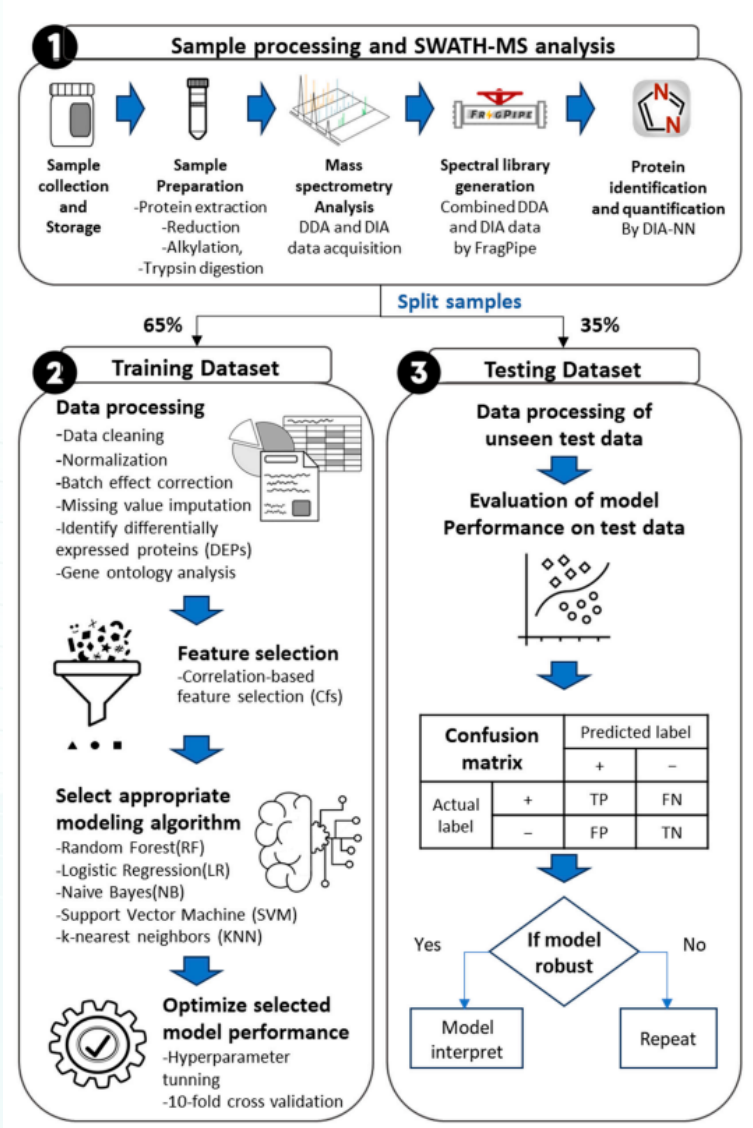
biomedicines

MDPI

Article

Application of SWATH Mass Spectrometry and Machine Learning in the Diagnosis of Inflammatory Bowel Disease Based on the Stool Proteome

Elmira Shajari [1,2,3], David Gagné [1,2,3,4], Mandy Malick [1,2,3], Patricia Roy [1,2,3], Jean-François Noël [4], Hugo Gagnon [4], Marie A. Brunet [2,5], Maxime Delisle [2,6], François-Michel Boisvert [2,3] and Jean-François Beaulieu [1,2,3,*]



Campioni:

- 123 campioni fecali da pazienti sintomatici (70 IBD, 53 non-IBD).
- Raccolta nel contesto di test per la **calprotectina fecale**.
- 78 campioni usati per training, 45 per testing (**blind set**).

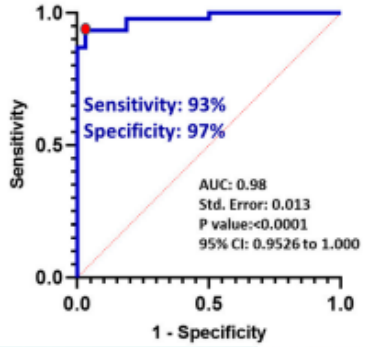
Valutati 5 algoritmi:

Support Vector Machine (SVM),
Random Forest (RF),
K-Nearest Neighbors (KNN),
Naive Bayes (NB),
Logistic regression (LR).

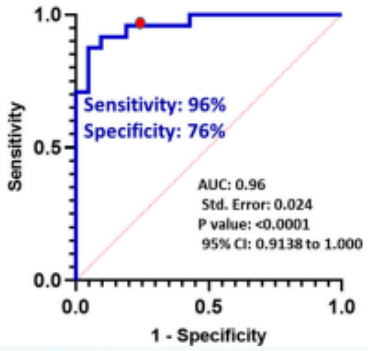
Support Vector Machine (SVM) con kernel lineare si è rivelato il migliore.

Classifier	Accuracy	Precision	Recall	F-Score	AU-ROC	AU-PRC
SVM	95%	0.97	0.93	0.96	0.95	0.96
NB	90%	0.94	0.90	0.92	0.93	0.94
LR	88%	0.89	0.91	0.90	0.92	0.92
KNN	88%	0.91	0.89	0.90	0.90	0.89
RF	87%	0.89	0.89	0.89	0.94	0.93

(a) ROC of optimized SVM model on training data



(b) ROC of evaluation of SVM model on testing data



c)

SVM prediction on test dataset		Predicted label	
		aIBD	Ctrl
Actual label	aIBD	23 (TP)	1 (FN)
	Ctrl	5 (FP)	16 (TN)

Risultati

- **Accuratezza sul training set** (10-fold CV): 95% (SVM).
- **Validazione su dati non visti (45 campioni)**:
  - **Sensibilità**: 96%
  - **Specificità**: 76%
  - **AUC**: 0.96

Conclusioni

- Il modello mostra grande **potenziale clinico** come test non invasivo per identificare fasi attive dell'IBD.
- Supera i limiti della sola calprotectina (che può dare falsi positivi/negativi).
- Potenzialmente utile per ridurre colonscopie inutili e migliorare la gestione terapeutica.

Punti di forza

- Approccio multidisciplinare: clinica, proteomica, bioinformatica.
- Valutazione rigorosa dell'intero flusso di analisi dati (normalizzazione, batch effect, imputazione).
- Validazione prospettica.

Limitazioni

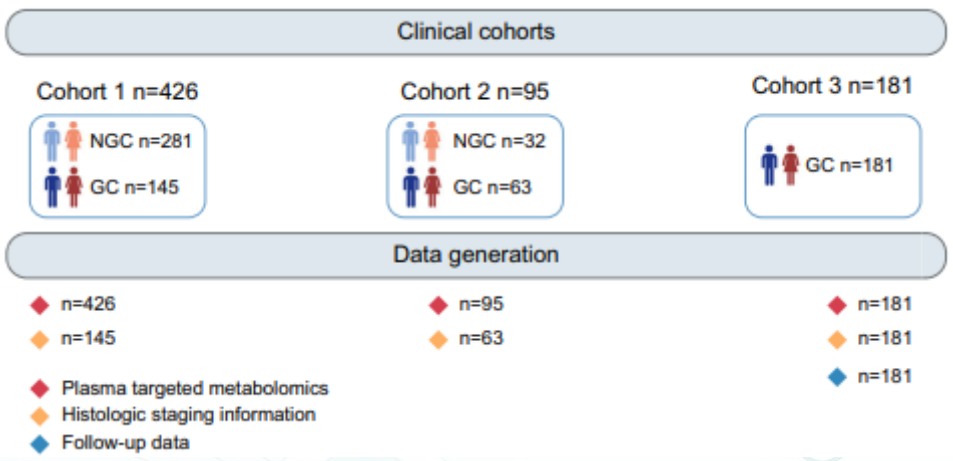
- Studio monocentrico, su adulti canadesi.
- Tutti i dati raccolti con un solo strumento MS (potenziale bias tecnico).
- Servono ulteriori validazioni multi-coorte e multi-strumentali.

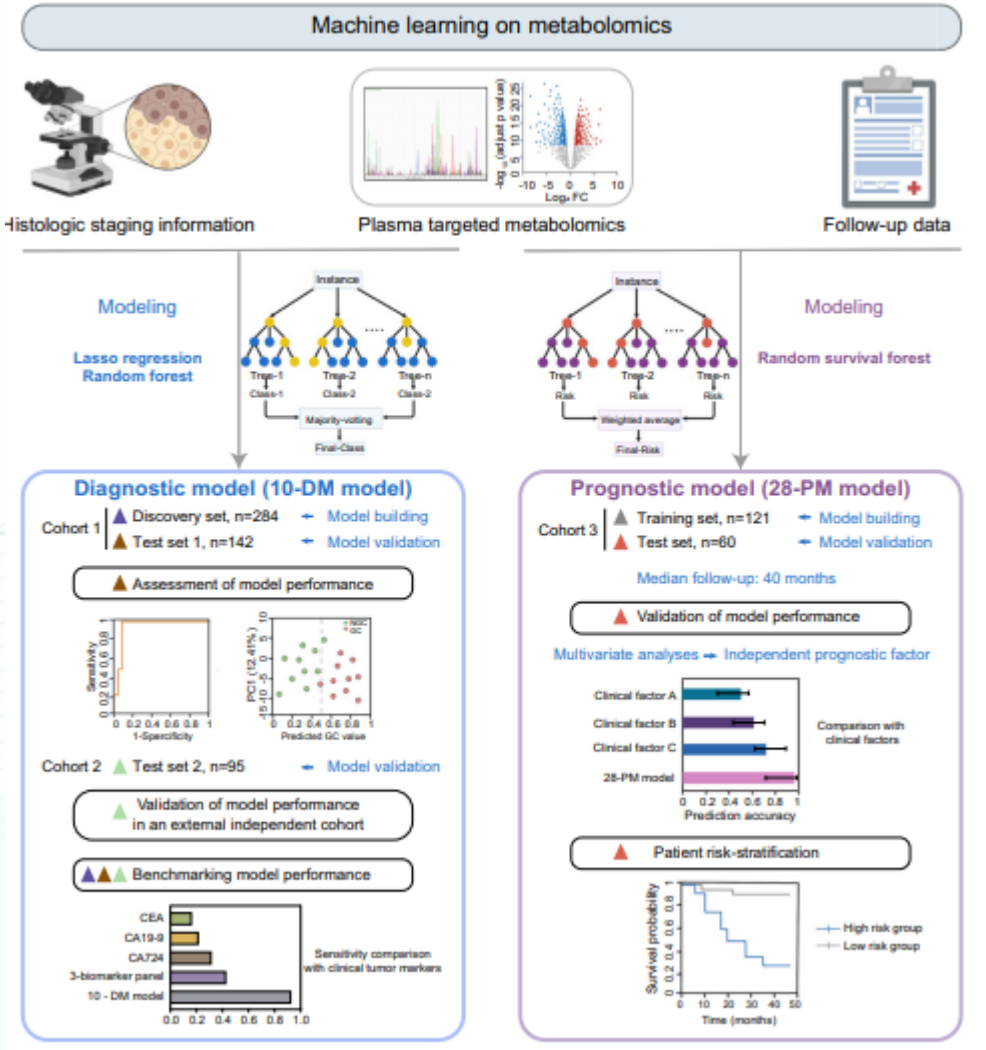
nature communications

Article

https://doi.org/10.1038/s41467-024-46043-y

Metabolomic machine learning predictor for diagnosis and prognosis of gastric cancer

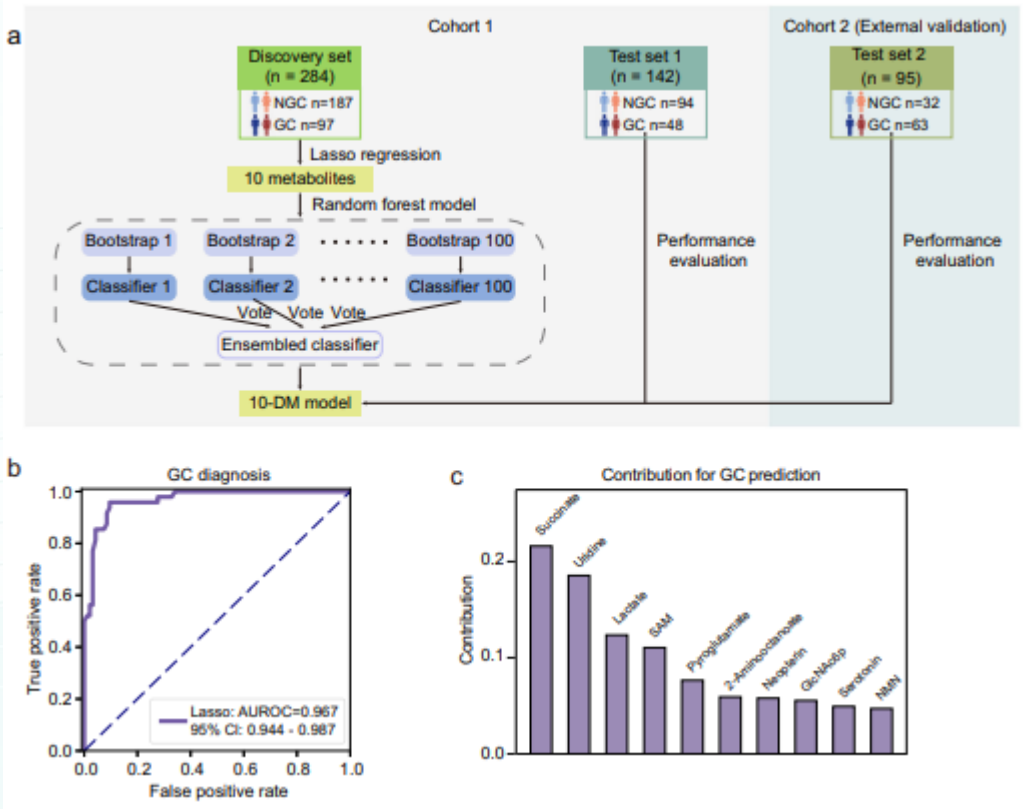




Coorti <u>multicentrici</u>:

- <u>Coorti 1 e 2:</u> per la costruzione e validazione del modello diagnostico.
- <u>Coorte 3:</u> per il modello prognostico con follow-up mediano di 40 mesi.

Sviluppare un modello capace di **diagnosticare il cancro gastrico in modo non invasivo**, con **alta sensibilità e specificità**, anche **nelle fasi iniziali** (stadio IA/IB).



1. LASSO (Least Absolute Shrinkage and Selection Operator)

•Tipo di **regressione lineare**.

•Serve per:

- **Selezionare le variabili rilevanti** (riduce a zero i coefficienti dei metaboliti meno informativi),
- **Evitare overfitting**.

•In questo studio: ha selezionato **10 metaboliti** da un set iniziale di 147.

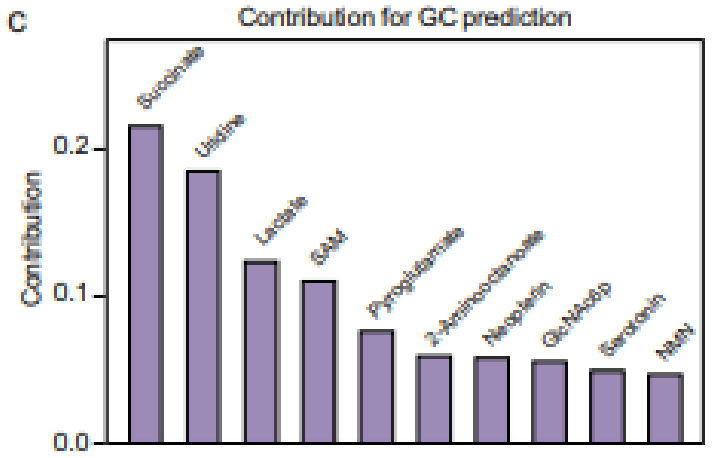
2. Random Forest

•Algoritmo **ensemble** basato su **molti alberi decisionali**.

•Ogni albero è allenato su un campione diverso e su sottoinsiemi casuali di metaboliti.

•Fa una **classificazione binaria** (GC vs NGC) usando il **voto di maggioranza** degli alberi.

•Fornisce una **probabilità di essere GC** (es. 0.8 → paziente classificato come GC).



↑ Aumentati in GC

Neopterina

Serotonina

SAM (S-Adenosyl Methionine)

NMN (Nicotinamide mononucleotide)

GlcNAc6P (N-Acetil-glucosamina-6-P)

↓ Ridotti in GC

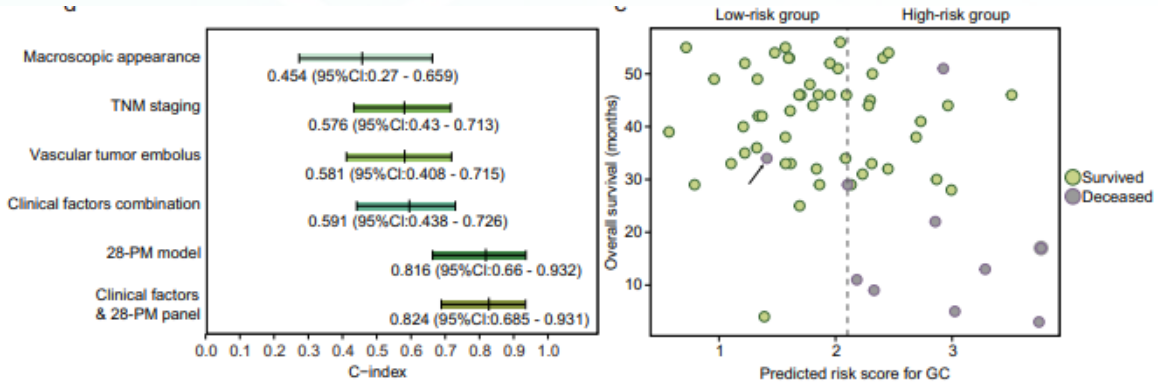
Succinato

Uridina

Lattato

Pyroglutammato

2-Aminoottanoato



Random Survival Forest (RSF)

È un algoritmo che:

- Costruisce **molti alberi decisionali** (forest),
- Ognuno è allenato su un **sottoinsieme casuale** dei dati,
- Ogni albero fa **stime di rischio o sopravvivenza** nel tempo,
- Le **previsioni finali** si ottengono **combinando le stime** di tutti gli alberi.

Prestazioni del modello

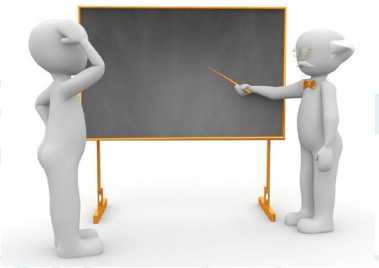
- **C-index** (indicatore della capacità predittiva): **0.83** sul test set (buon valore).
- Prestazioni **migliori** rispetto ai tradizionali fattori clinici come stadio TNM o morfologia macroscopica del tumore.
- Utilizzabile per **guidare la medicina di precisione**, cioè decidere:
  - Chi ha bisogno di terapie più aggressive,
  - Chi può essere seguito con minor intensità.

Cosa fa il modello

- Calcola un **"risk score"** per ciascun paziente.
- In base al punteggio, i pazienti sono **stratificati in gruppi a basso o alto rischio**.
- I pazienti **ad alto rischio** hanno:
  - Maggiore probabilità di morire o avere recidive.
  - Peggior sopravvivenza globale (OS) e sopravvivenza libera da malattia (DFS).

Conclusioni

- La **diagnosi tradizionale** ha limiti: soggettività, bassa sensibilità, focus riduttivo
- La **spettrometria di massa** permette di esplorare in profondità il metabolismo
- Il **machine learning** interpreta dati complessi, individua pattern nascosti
- L'integrazione dei due crea modelli **diagnostici e predittivi** ad alta accuratezza



Prospettive future

- Verso una **medicina di precisione** basata su dati molecolari
- Integrazione con genomica, microbioma, immagini e cartelle cliniche
- Sfide: validazione clinica, standardizzazione, etica nell'uso dell'IA