Project Report

Marco Savastano
Carmine Vardaro

Information Engineering for Digital Medicine
Artificial Intelligence for Omics Data Analysis Course 2025-2026

Contents

Abstract

The short abstract (50-80 words) is intended to give the reader an overview of the work.

1 Introduction

1.1 Background Clinico

Breve panoramica sulla patologia (CHD) e importanza di trovare nuovi biomarcatori non invasivi.

1.2 La Metabolomica Untargeted

Perché la LC-MS è la scelta giusta qui (visione olistica del fenotipo).

1.3 Problematiche Aperte

Qui introduci il ”problema” del tuo progetto: la complessità dei dati, la necessità di integrare due modi di ionizzazione (ESI+/ESI-) e la scelta del miglior metodo di preprocessing (non esiste una ”ricetta unica”).

1.4 Scopo del Lavoro

Valutare e ottimizzare un workflow chemiometrico (dal preprocessing alla Data Fusion) per distinguere soggetti Sani vs Patologici e identificare le feature biologicamente rilevanti.

2 Materials and Methods

2.1 Descrizione del Dataset

The dataset is composed of metabolomic data acquired by LC–MS in both positive (ESI+) and negative (ESI-) ionisation mode.

There are no missing values and zeros, due to an value inputation phase already done focused on the replacement with one-fifth of the minimum value recorded in the dataset for that molecule. [1]

The following tables give an overview on the dataset composition, with the corresponding values:

TABLE I: ESI- Dataset Distribution and Characteristics

DESCRIPTION	VALUE
Total Samples	219
Total Features (Metabolites)	52
Class Count: CTRL	107
Class Count: CHD	104
Class Count: QC	8
Samples with suffix ’_00’	28
Samples with suffix ’_01’ (Tech Replicate)	14
Samples without suffix	177
Estimated Unique Biological Samples	205
CTRL - Biological Samples	100
CTRL - Technical Replicates	7
CHD - Biological Samples	97
CHD - Technical Replicates	7
QC - Total Samples	8
Negative Values Present	No

TABLE II: ESI+ Dataset Distribution and Characteristics

DESCRIPTION	VALUE
Total Samples	219
Total Features (Metabolites)	98
Class Count: CTRL	107
Class Count: CHD	104
Class Count: QC	8
Samples with suffix '_00'	28
Samples with suffix '_01' (Tech Replicate)	14
Samples without suffix	177
Estimated Unique Biological Samples	205
CTRL - Biological Samples	100
CTRL - Technical Replicates	7
CHD - Biological Samples	97
CHD - Technical Replicates	7
QC - Total Samples	8
Negative Values Present	No

2.2 Quality Assessment & Data Cleaning

Descrizione dei QC e dei Replicati Tecnici. **Nota:** Spiegare chiaramente che QC e duplicati sono stati utilizzati per una valutazione iniziale della stabilità ma rimossi dal dataset finale di modeling (es. numero esiguo, impossibilità di mediare).

2.3 Strategie di Pre-processing (Iterativo)

Elenco delle tecniche testate: Normalizzazione, Trasformazione Logaritmica, Scaling (Autoscaling).

2.4 Anomaly Detection

Descrizione del metodo usato per identificare e rimuovere gli outlier (fondamentale per la pulizia del dato).

2.5 Feature Selection

Metodi statistici/algoritmici utilizzati per ridurre la dimensionalità e rimuovere il rumore prima del modeling.

2.6 Strategie di Data Fusion

Definizione degli approcci: Low-Level (concatenazione semplice) vs High-Level (o altri approcci). L'obiettivo è sfruttare la complementarità ESI+/ESI-.

2.7 Analisi Statistica e Machine Learning

- **Unsupervised:** PCA (per l'esplorazione).
- **Supervised:** PLS-DA, SVM, Random Forest, Logistic Regression.
- **Validazione:** Descrizione rigorosa dello split Training Set vs Validation Set (o Cross-Validation) per evitare l'overfitting.

2.8 Stack Tecnologico

Breve paragrafo sulle librerie Python utilizzate (Pandas, Scikit-learn, ecc.) per garantire la riproducibilità.

3 Results and Discussion

3.1 Valutazione dell'Analisi Esplorativa (PCA)

Visualizzazione dei dati ESI+ e ESI- separati. Valutazione degli Outlier (prima e dopo la rimozione). Confronto delle tecniche di scaling (es. efficacia dell'autoscaling).

3.2 Performance dei Modelli su Singoli Dataset (ESI+ / ESI-)

Confronto delle metriche (Accuratezza, Specificità, Sensibilità) tra PLS-DA, SVM, RF, LR. Quale modello performa meglio sui dati positivi? E sui negativi?

3.3 Risultati della Data Fusion

La fusione dei dati ha migliorato la classificazione rispetto ai dataset singoli?

3.4 Interpretabilità e Biomarcatori (Feature Importance)

Analisi delle Feature Importances e Analisi Univariata (Volcano Plot). Identificazione/interpretazione biologica dei top-metaboliti.

4 Conclusions

Sintesi del miglior workflow identificato. Considerazioni sull'interpretabilità biologica e limiti dello studio (es. numero di campioni, assenza di validazione esterna).

References

[1] Mires, Stuart, et al. "Plasma metabolomic and lipidomic profiles accurately classify mothers of children with congenital heart disease: an observational study." Metabolomics 20.4 (2024): 70.

TABLE III: Simulation Parameters

Information message length	$k = 16000$ bit
Radio segment size	$b = 160$ bit
Rate of component codes	$R_{cc} = 1/3$
Polynomial of component encoders	$[1, 33/37, 25/37]_8$

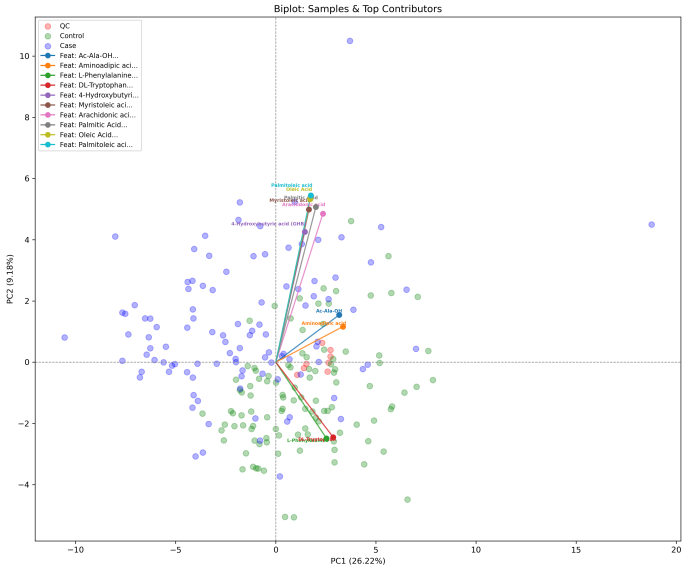

Fig. 1: Simulation results on the AWGN channel. Average throughput k/n vs E_s/N_0.