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

Origine dei dati (LC-MS). I due blocchi: ESI+ e ESI-. Composizione delle classi (Controlli vs CHD).

2.2 Quality Assessment and Data Cleaning

This preliminary phase is crucial to validate the technical quality of the experiment before proceeding with biological interpretation. The objectives of this assessment were twofold: (i) to verify the instrumental stability over the analytical run, and (ii) to evaluate the technical reproducibility of the measurements.

To this end, we monitored the behavior of Quality Control (QC) samples (pooled aliquots injected periodically) and technical replicates (samples analyzed in duplicate, denoted with suffixes _00 and _01). The evaluation was performed using Principal Component Analysis (PCA) applied separately to the raw data of Negative (ESI-) and Positive (ESI+) ionization modes. Autoscaling was applied prior to PCA to ensure all metabolites contributed equally to the model, regardless of their absolute intensity.

2.2.1 PCA on Raw Data: Unsupervised Evaluation

PCA was employed as an unsupervised exploratory technique to visualize the intrinsic structure of the data variance. To facilitate the inspection of instrumental stability and technical reproducibility, specific sub-plots were generated from the same global PCA model to isolate QC samples and technical replicates. The results for the Negative and Positive ionization modes are presented in Figure 1 and Figure 2, respectively.

a) Variance and Model Structure

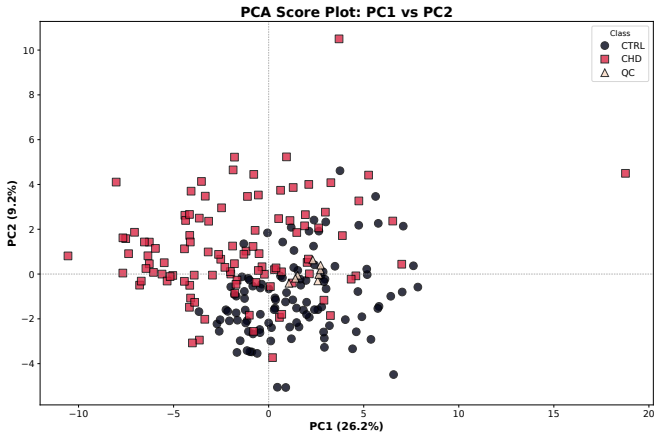
The Scree Plots (Fig. 1c, 2c) display the percentage of variance explained by each Principal Component (PC). The Loadings Plots (Fig. 1b, 2b) and Profiles (Fig. 1d-e, 2d-e) provide an overview of the features driving the separation.

b) Instrumental Stability (QC Analysis)

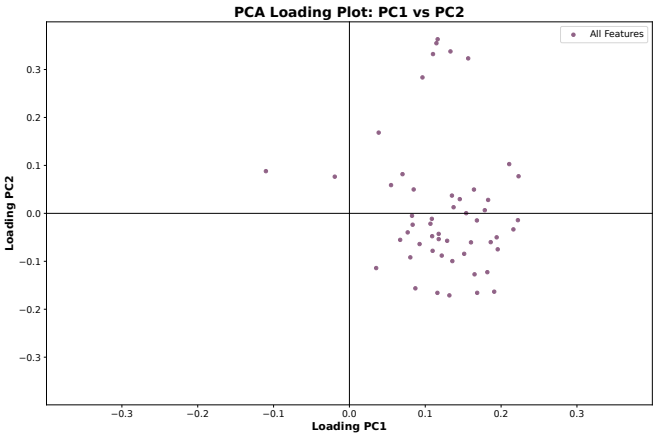
The stability of the LC-MS system was evaluated by isolating the QC samples in the PCA space. As shown in the QC-specific score plots (Fig. 1f, 2f), only the QC samples are visualized to assess their compactness. In the **Negative Ionization mode (ESI-)**, the QCs form a tight, well-defined cluster, indicating high instrumental stability. In the **Positive Ionization mode (ESI+)**, the QCs show a slightly higher dispersion. This behavior is attributed to the inherent characteristics of Hydrophilic Interaction Liquid Chromatography (HILIC). The HILIC separation mechanism relies on a water-enriched layer on the stationary phase, making the partitioning equilibrium more sensitive to minor fluctuations in column conditioning compared to reversed-phase chromatography. However, despite this inherent dispersion, the QC cluster remains distinct from the biological variability.

c) Technical Reproducibility (Replicates Analysis)

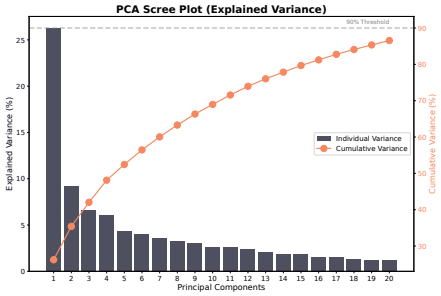
Reproducibility was assessed by projecting the technical replicates onto the PCA space (Fig. 1g, 2g). In these plots, non-replicated samples are hidden to highlight the distance between paired measurements (_00 and _01). In both ionization modes, the pairs of replicates are projected in close proximity, often overlapping, confirming that the analytical workflow yields consistent results for the same biological sample.



(a) Global Score Plot (All Samples)

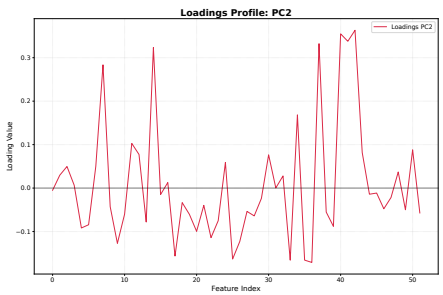


(b) Global Loading Plot

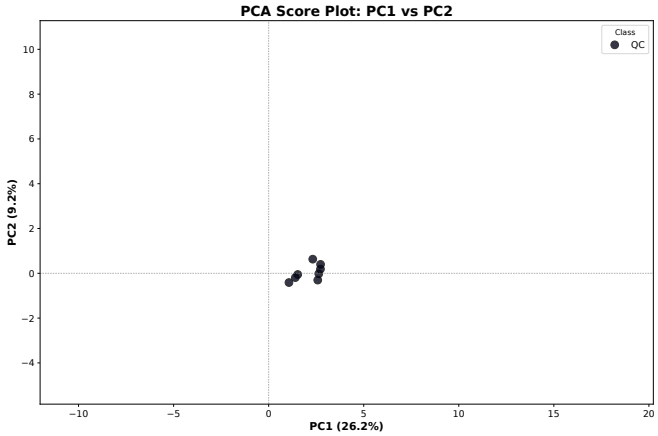


(c) Scree Plot

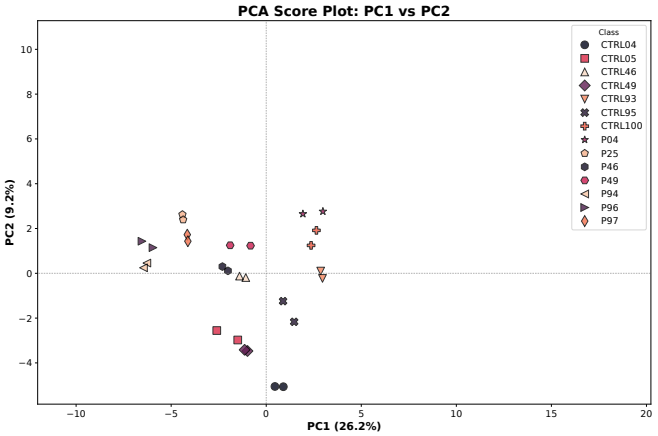
(d) Loading Profile PC1



(e) Loading Profile PC2



(f) Instrumental Stability (QC only)



(g) Reproducibility (Replicates only)

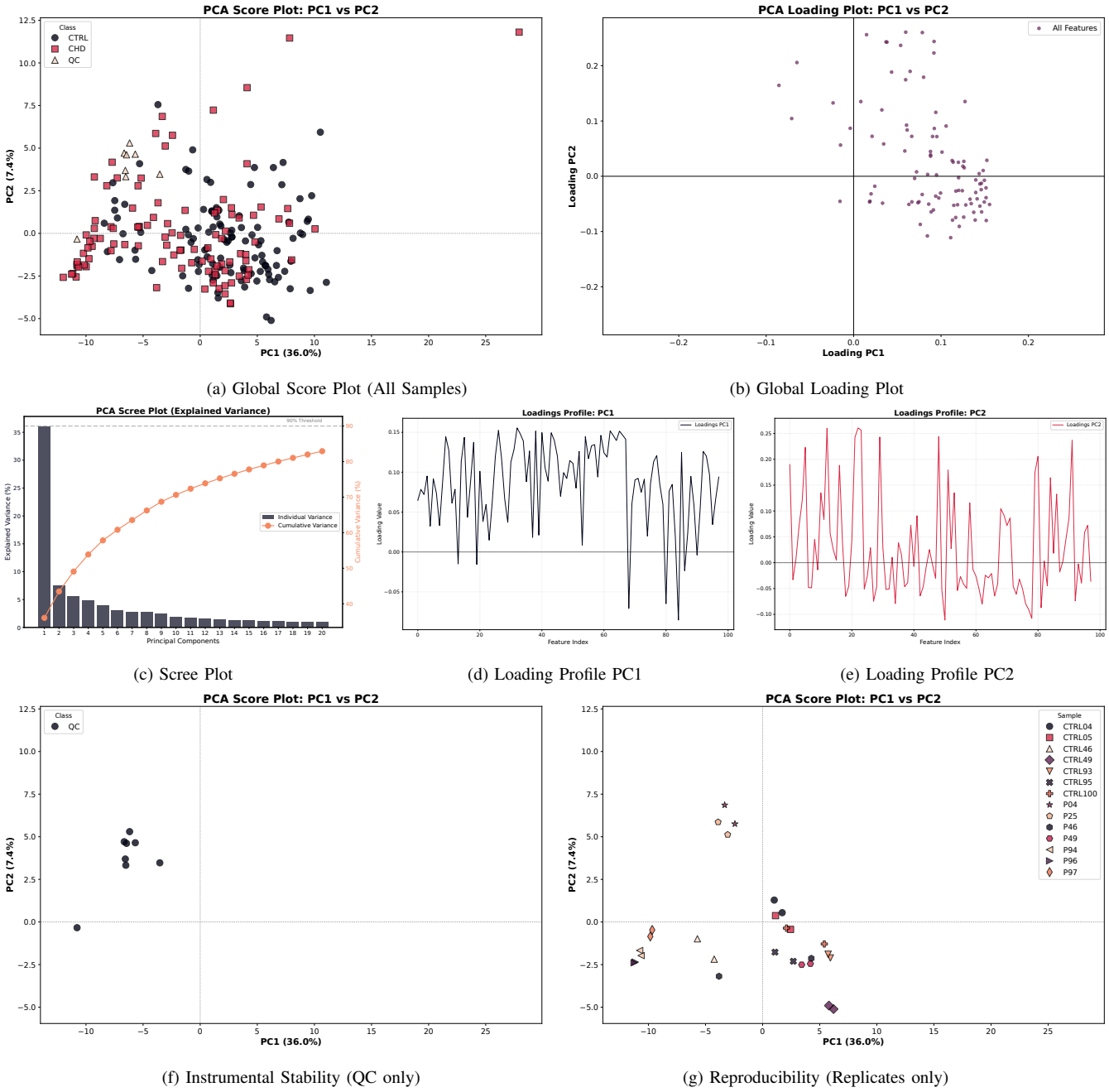
Fig. 1: **Quality Assessment for ESI- Dataset.** (a) Global PCA Score plot showing the distribution of all classes (CHD, CTRL, QC). (b) Loading plot showing feature contributions. (c-e) Variance analysis and loading profiles. (f) Zoom on QC samples: the tight cluster confirms high stability. (g) Zoom on technical replicates: paired samples show high overlap, confirming reproducibility.

2.2.2 Dataset Refinement for Downstream Analysis

Having confirmed the technical robustness of the experiment, specific data cleaning steps were implemented to prepare the dataset for biological modeling.

a) Removal of Quality Control Samples

QC samples were removed from the final dataset as they have fulfilled their purpose of monitoring instrumental stability. Retaining QCs in downstream supervised analysis (e.g., PLS-DA) would introduce an artificial class that does not reflect a biological phenotype. Furthermore, due to their chemical homogeneity, QCs would form a dense cluster accounting for a large portion of the total variance, potentially masking the subtler biological differences between CHD and CTRL groups.



(a) Global Score Plot (All Samples)

(b) Global Loading Plot

(c) Scree Plot

(d) Loading Profile PC1

(e) Loading Profile PC2

(f) Instrumental Stability (QC only)

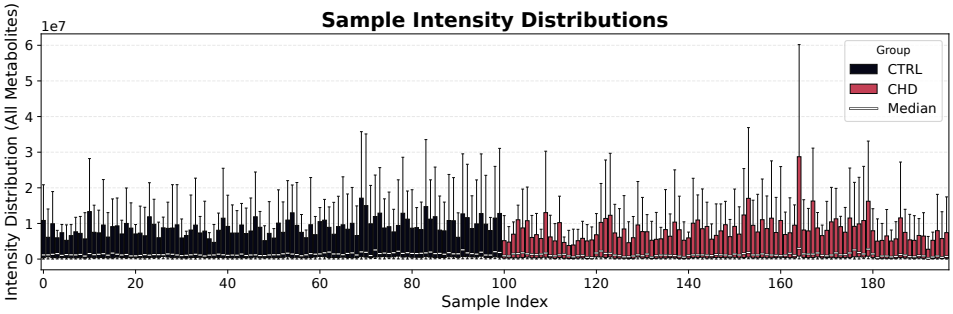
(g) Reproducibility (Replicates only)

Fig. 2: **Quality Assessment for ESI+ Dataset.** (a-b) Global PCA model overview. (c-e) Variance and Loadings. (f) Stability check: QCs show a wider dispersion compared to negative mode, consistent with HILIC sensitivity, but remain distinct from biological variance. (g) Reproducibility check showing paired replicates.

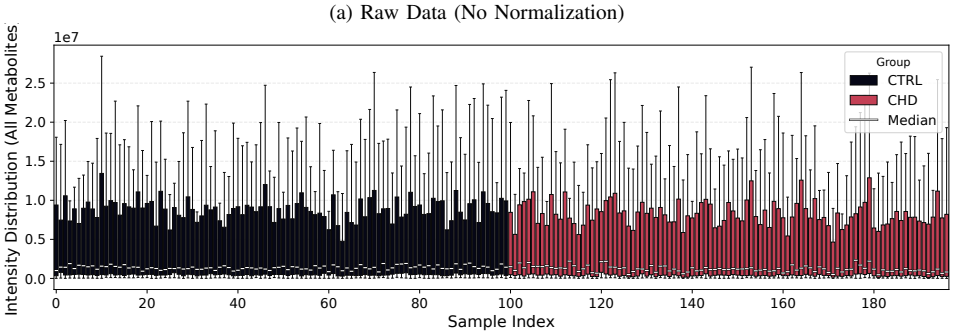
b) Removal of Technical Duplicates

To ensure the statistical independence of observations, technical duplicates were handled by retaining only one measurement per biological subject (samples with suffix _00). Including both replicates would violate the assumption of independence required by most statistical tests, artificially inflating the sample size and underestimating the intra-class variance. As duplicates were not available for all samples, averaging was avoided to prevent inconsistency in the data structure. Therefore, the removal of the second replicate (_01) ensures a homogeneous dataset where each sample represents a unique biological entity.

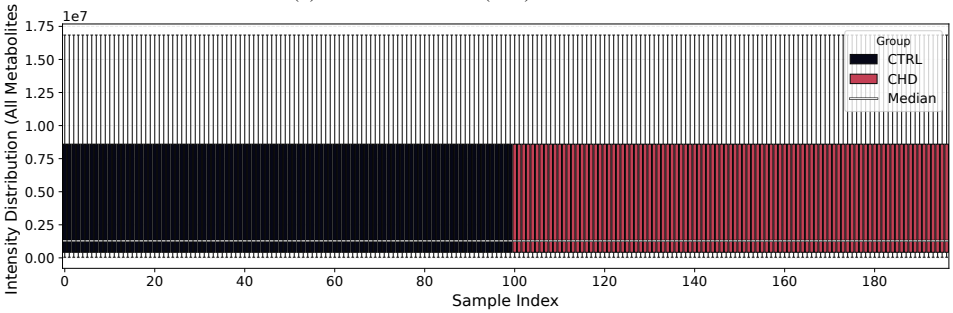
2.3 Strategie di Pre-processing



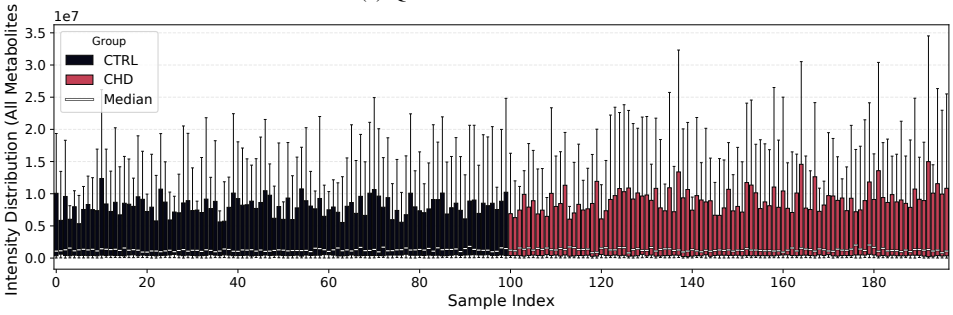
(a) Raw Data (No Normalization)



(b) Total Ion Current (TIC) Normalization



(c) Quantile Normalization



(d) Probabilistic Quotient Normalization (PQN)

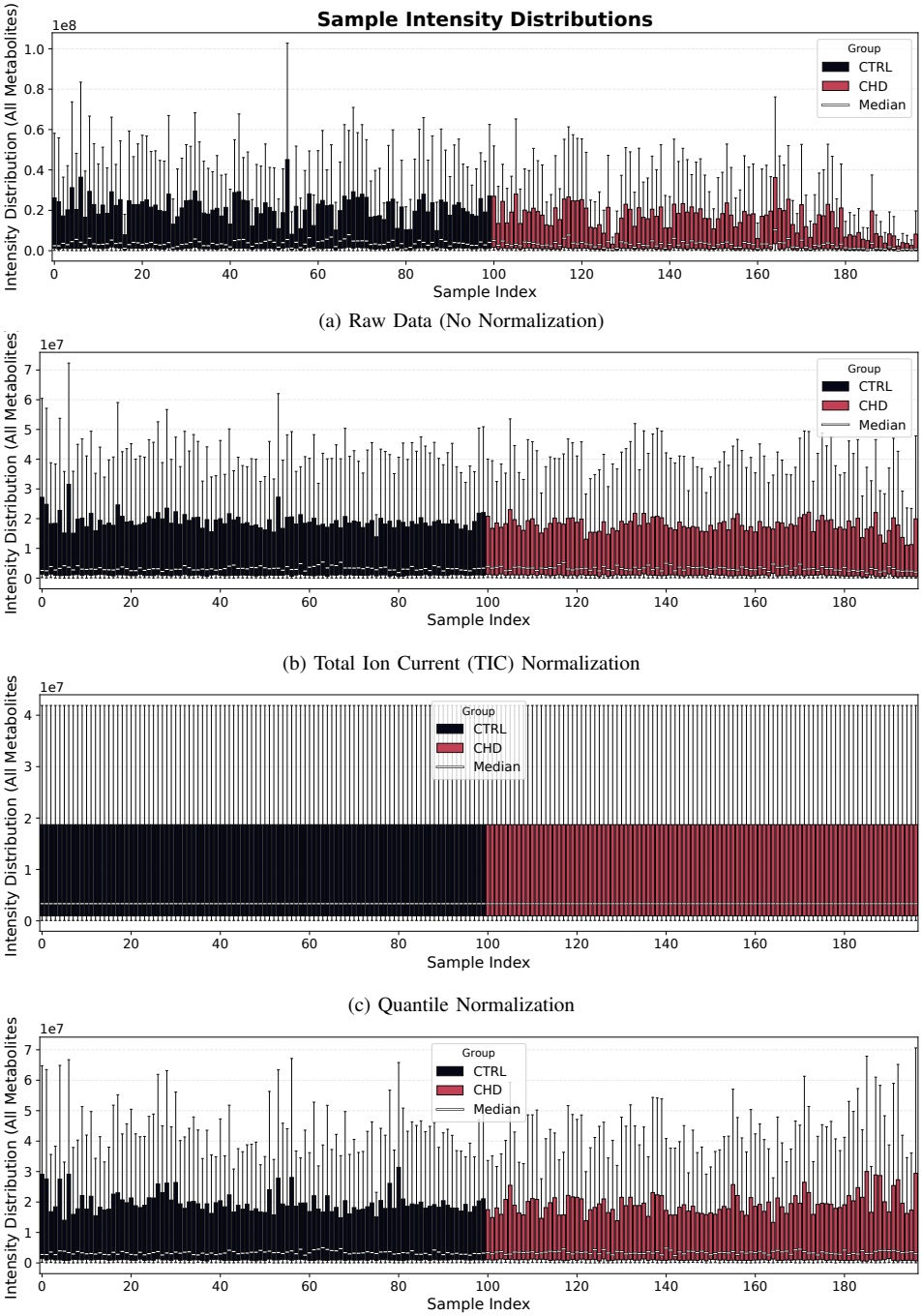
Fig. 3: **Comparison of Normalization Strategies for ESI- Dataset.** Boxplots representing the global intensity distribution of all samples. (a) Raw data showing significant systematic variation (e.g., dilution effects). (b) TIC normalization, acting on the total sum. (c) Quantile normalization, forcing identical distributions potentially suppressing biological signal. (d) PQN, the selected method, which effectively reduces technical variance while preserving biological information.

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.



(a) Raw Data (No Normalization)

(b) Total Ion Current (TIC) Normalization

(c) Quantile Normalization

(d) Probabilistic Quotient Normalization (PQN)

Fig. 4: **Comparison of Normalization Strategies for ESI+ Dataset.** (a) Raw data distribution. (b) TIC normalization results. (c) Quantile normalization results showing aggressive distribution alignment. (d) PQN results, selected as the optimal compromise for downstream analysis.

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] J. Hagenauer, E. Offer, and L. Papke. Iterative decoding of binary block and convolutional codes. *IEEE Trans. Inform. Theory*, vol. 42, no. 2, pp. 429—445, Mar. 1996.
[2] T. Mayer, H. Jenkac, and J. Hagenauer. Turbo base-station cooperation for intercell interference cancellation. *IEEE Int. Conf. Commun. (ICC)*, Istanbul, Turkey, pp. 356–361, June 2006.
[3] J. G. Proakis. *Digital Communications*. McGraw-Hill Book Co., New York, USA, 3rd edition, 1995.
[4] F. R. Kschischang. Giving a talk: Guidelines for the Preparation and Presentation of Technical Seminars. http://www.comm.toronto.edu/frank/guide/guide.pdf.
[5] IEEE Transactions LaTeXand Microsoft Word Style Files. http://www.ieee.org/web/publications/authors/transjnl/index.html

TABLE I: 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	$[1, 33/37, 25/37]_8$