

Project Report

Marco Savastano
Carmine Vardaro

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

Contents

1	Introduction	2
1.1	Clinical background	2
1.2	Untargeted metabolomic	2
1.3	Open problematics	2
1.4	Project goal	2
2	Materials and Methods	2
2.1	Dataset description	2
2.2	Quality Assessment and Data Cleaning	3
2.3	Data Pre-Processing	6
2.3.1	Data Normalization	6
2.3.2	Data Transformation	8
2.3.3	Data Scaling	9
2.4	Anomaly Detection	11
2.5	Feature Selection	11
2.6	Data Fusion strategies	11
2.7	Statistical analysis and Machine Learning	11
2.8	Technology stack	11
2.8.1	NumPy	11
2.8.2	Pandas	11
2.8.3	Scikit-learn	12
2.8.4	Statsmodels	12
2.8.5	SciPy	12
2.8.6	Matplotlib	12
2.8.7	Seaborn	12
2.8.8	tqdm	12
2.8.9	os	12
3	Results and Discussion	12
3.1	Valutazione dell'Analisi Esplorativa (PCA)	12
3.2	Performance dei Modelli su Singoli Dataset (ESI+ / ESI-)	12
3.3	Risultati della Data Fusion	12
3.4	Interpretabilità e Biomarcatori (Feature Importance)	12
4	Conclusions	12
4.1	Future work	12
References		12

Abstract

The project is mainly focused on using LC-MS plasma metabolomics to address the identification of non-invasive maternal biomarkers for Congenital Heart Disease (CHD). The core was optimizing a robust workflow, by systematically evaluating pre-processing strategies, to prepare the dataset for predictive models, in pursuance of the research of metabolic alteration associated with fetal CHD risk.

1 Introduction

1.1 Clinical background

The Congenital Heart Disease (CHD) corresponds to the most prevalent class of congenital anomalies, representing in pediatric morbidity and mortality a significant burden. Current prenatal diagnostic methods are mostly operator-dependant and may be faulty during early stages of gestation. Therefore, the importance in finding a non-invasive biomarker material which can facilitate early risks assessments, is critical in view of improving screening accuracy. The analysis of maternal plasma, can result and offer an interesting window into the complex biochemical environment of the growing fetus, in revealing of the metabolic disregulation associated with CHD pathogenesis which can exceeds solely genetic factors.

1.2 Untargeted metabolomic

A powerful analytical approach for phenotype characterization is the untargeted metabolomic, able to acquire the downstream end-products of genomic, transcriptomic, and proteomic processes. Opposing to target assays that quantify a limited list of pre-defined compounds, the main intent of untargeted Liquid Chromatography-Mass Spectrometry (LC-MS) is to ensure an exhaustive profile of small molecules with biological sample in a hypothesis-free manner. From an holistic perspective, this technique can result in becoming a considerable advantage in the investigation of complex pathologies such as CHD, where some more specific metabolic pathways may not be fully understood. By obtaining data in both positive (ESI+) and negative (ESI-) ionization modes, there is the capability of an unbiased detection of a diverse range of chemical classes and providing a detailed snapshot of the maternal metabolic phenotype (metabotype).

1.3 Open problematics

The high dimensionality and complexity of biological variance of the output of the LC-MS data places a significant computational challenge for the analysis. The primary notable issue is the absence of a gold standard for data pre-processing, critically influence the validity of downstream biological conclusions. Sometimes however, this absence can result in a more advantageous scenario, allowing for more accurate and tailored data preparation in view of the problem at hand. Systematic technical variations, such as sample dilution effects or instrumental fluctuations, has to be revised without removing genuine biological heterogeneity. Furthermore, the dataset composition and structure necessitate effective strategies, remining a complex analytical hurdle addressed in this project.

1.4 Project goal

The primary objective of this work is to delineate, evaluate and optimize a comprehensive chemiometric analysis workflow of the metabolomic plasma, in the context of CHD. One of the project intent was systematically asses and comparing the impact of different pre-processing strategies, to determine the best strategy for minimizing technical variance while preserving biological signal. Additionally, the application of Data Fusion techniques was utilized to integrate ESI+ and ESI- modalities, thereby exploiting their complementarity. Ultimately, the machine learning models were employed to determine and discriminate between healthy and pathological pregnancies and to identify statistically significant metabolic features that may serve as potential clinical biomarkers.

2 Materials and Methods

2.1 Dataset description

The dataset is composed of plasma metabolomic data derived from a cohort of pregnant women and acquired using Liquid Chromatography-Mass Spectrometry (LC-MS). The analysis was performed in both positive (ESI+) and negative (ESI-) ionisation mode, to guarantee a sufficient metabolic coverage. The dataset is divided into different distinct classes: incl

- CTRL (Control) samples related to mother of healthy children
- CHD for Congenital Heart Disease pathologic cases
- QC (Quality Control) samples for each ionization mode, prepared by pooling equal fractions of all biological samples injected periodically to monitor instrumental stability and ensure analytical reproducibility.

The high-dimensionality data blocks are composed with metabolites on each row and individual biological samples on the columns. There are no missing values and zeros, due to a value imputation 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, for both ESI+ and ESI- 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 and Data Cleaning

This preliminary phase is crucial to validate the technical quality of the experiment before proceeding with biological interpretation.

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.

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.

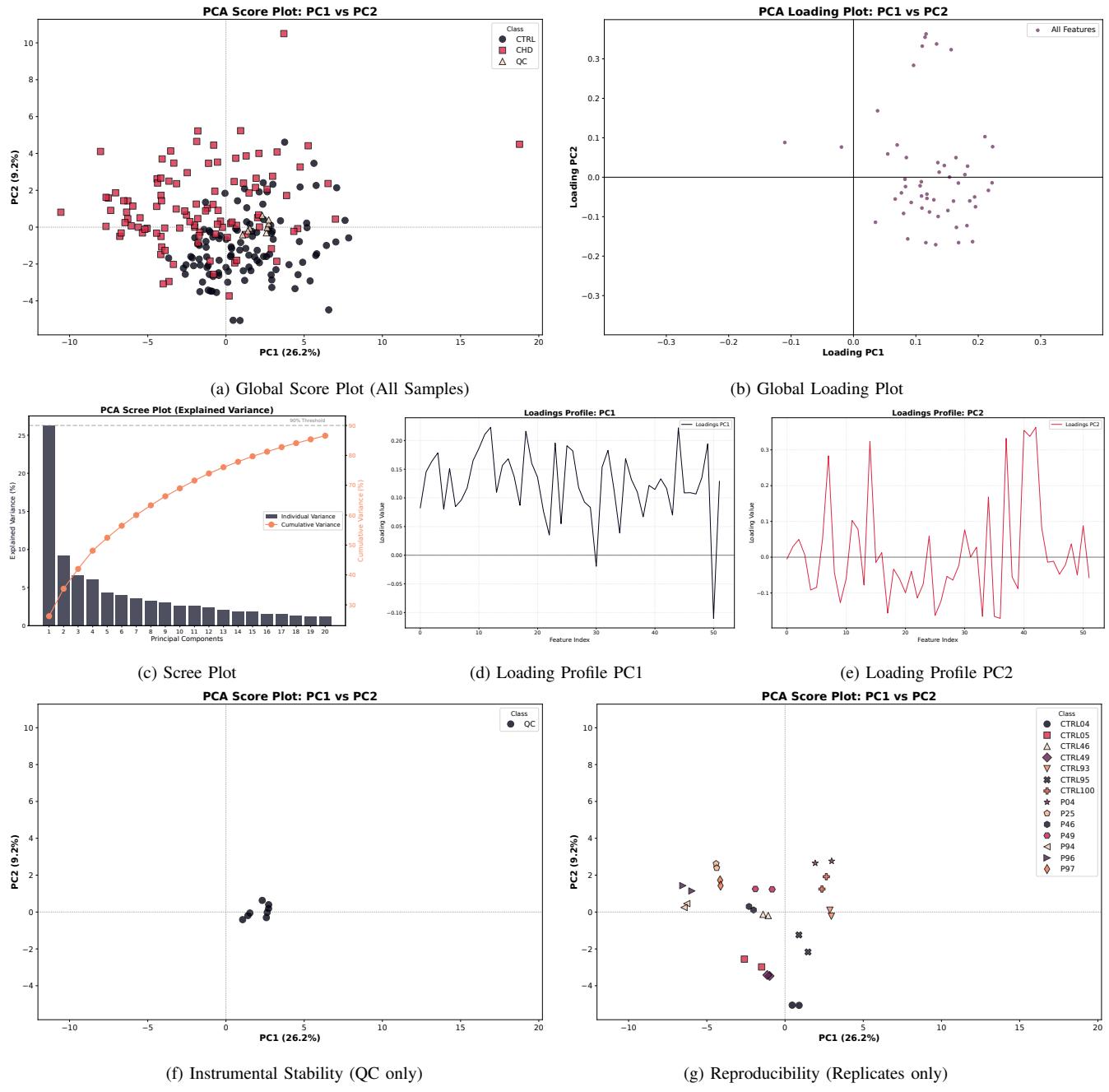


Fig. 1: **Quality Assessment for ESI-Dataset.** (a) Global PCA Score plot showing the distribution of all classes (CHD, CTRL, QC). (b) Loading plot showing feature contributions. (c-e) Variance analysis and loading profiles. (f) Zoom on QC samples: the tight cluster confirms high stability. (g) Zoom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

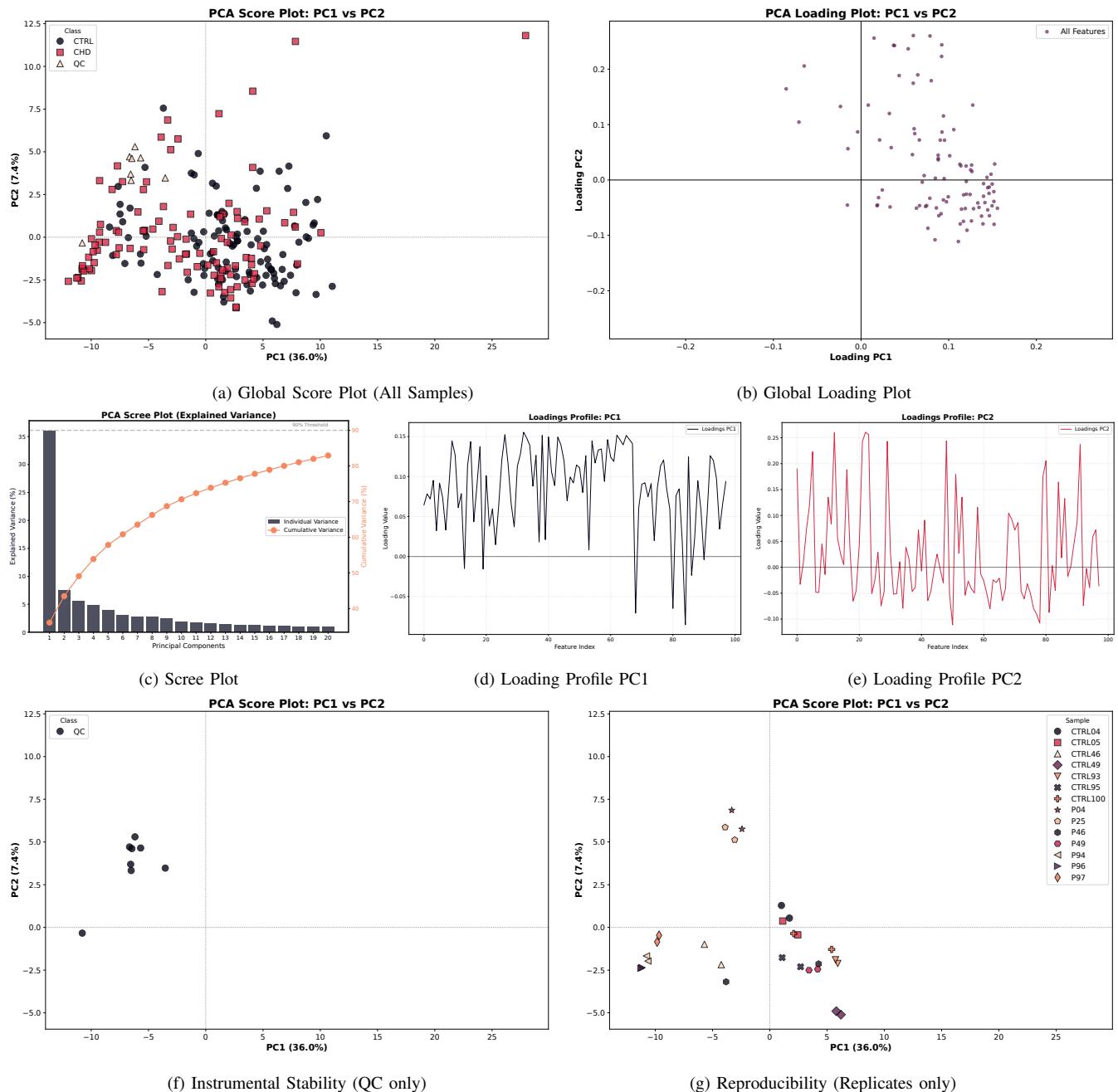


Fig. 2: **Quality Assessment for ESI+ Dataset.** (a-b) Global PCA model overview. (c-e) Variance and Loadings. (f) Stability check: QCs show a wider dispersion compared to negative mode, consistent with HILIC sensitivity, but remain distinct from biological variance. (g) Reproducibility check showing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

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

d) 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.

e) 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 Data Pre-Processing

2.3.1 Data Normalization

The LC-MS untargeted analysis of biological fluids is inherently subject to systematic variations unrelated to the biological problem, such as differences in sample dilution (e.g., hydration status of the subjects) and fluctuations in ionization efficiency. As evidenced by the raw data distribution (Fig. 3a and Fig. 4a), a significant variability in the median intensity across samples was observed, necessitating a normalization step to render the samples comparable.

We evaluated multiple normalization strategies, ranging from global intensity corrections (TIC, Mean, Median) to distribution-based methods (Quantile) and robust probabilistic approaches (Probabilistic Quotient Normalization - PQN). The selection of the optimal method was driven by a dual criterion: (i) qualitative inspection of sample distributions via boxplots, and (ii) quantitative assessment of the Coefficient of Variation (CV%) calculated on the technical replicates and across the biological groups.

Figures 3 and 4 illustrate the effect of selected normalization algorithms on the Negative (ESI-) and Positive (ESI+) datasets, respectively. While the raw data showed pronounced "batch-like" or dilution-related fluctuations, all normalization methods improved the alignment of sample medians. Specifically:

- **Total Ion Current (TIC):** Provided a standard correction based on the total signal sum, effectively reducing global differences but potentially sensitive to high-intensity artifacts.
- **Quantile Normalization:** Resulted in perfectly aligned distributions (Fig. 3c, 4c). However, visual inspection suggests this approach may be overly aggressive, forcing all samples to conform to an identical distribution and potentially suppressing genuine biological heterogeneity.
- **Probabilistic Quotient Normalization (PQN):** Demonstrated a robust alignment of medians and interquartile ranges (Fig. 3d, 4d) without imposing the artificial uniformity observed with Quantile normalization.

To objectively quantify the reduction in technical variance, the median Coefficient of Variation (CV%) was calculated for all features across the Control (CTRL) and Disease (CHD) groups. The results are summarized in Table III and Table IV.

In both ionization modes, the absence of normalization yielded the highest variability (Avg CV ≈ 58% for ESI- and 68% for ESI+). Consistent with the visual inspection, **Quantile normalization** achieved the lowest numerical CV values (Avg CV ≈ 50% and 57%, respectively). However, minimal variance is not solely indicative of data quality; it may also reflect overfitting and loss of biological signal. **PQN** consistently ranked among the top performing methods, achieving a substantial reduction in variance (Avg CV ≈ 51.5% for ESI- and 59.7% for ESI+) comparable to Quantile and Median normalization, while theoretically preserving the relative abundance ratios of metabolites better than global sum methods.

Based on the combined evidence, **PQN (Probabilistic Quotient Normalization)** was selected as the optimal strategy for this study. It provides the best trade-off between the reduction of systematic error (comparable to the most aggressive methods) and the preservation of biological variance required for the subsequent biomarker discovery phase.

TABLE III: Comparison of Normalization Methods by Coefficient of Variation (CV%) - ESI Negative Dataset.

Normalization Method	Median CV CTRL (%)	Median CV CHD (%)	Average CV (%)
Quantile	45.46	55.10	50.28
PQN	44.94	58.06	51.50
Mean / TIC	45.91	59.66	52.79
Median	45.48	63.94	54.71
Range / Max	51.28	63.22	57.25
None (Raw)	49.11	67.40	58.26

TABLE IV: Comparison of Normalization Methods by Coefficient of Variation (CV%) - ESI Positive Dataset.

Normalization Method	Median CV CTRL (%)	Median CV CHD (%)	Average CV (%)
Quantile	53.35	61.62	57.48
Median	54.48	60.69	57.58
Range / Max	53.84	64.46	59.15
PQN	56.12	63.38	59.75
Mean / TIC	54.44	65.09	59.77
None (Raw)	58.43	77.84	68.14

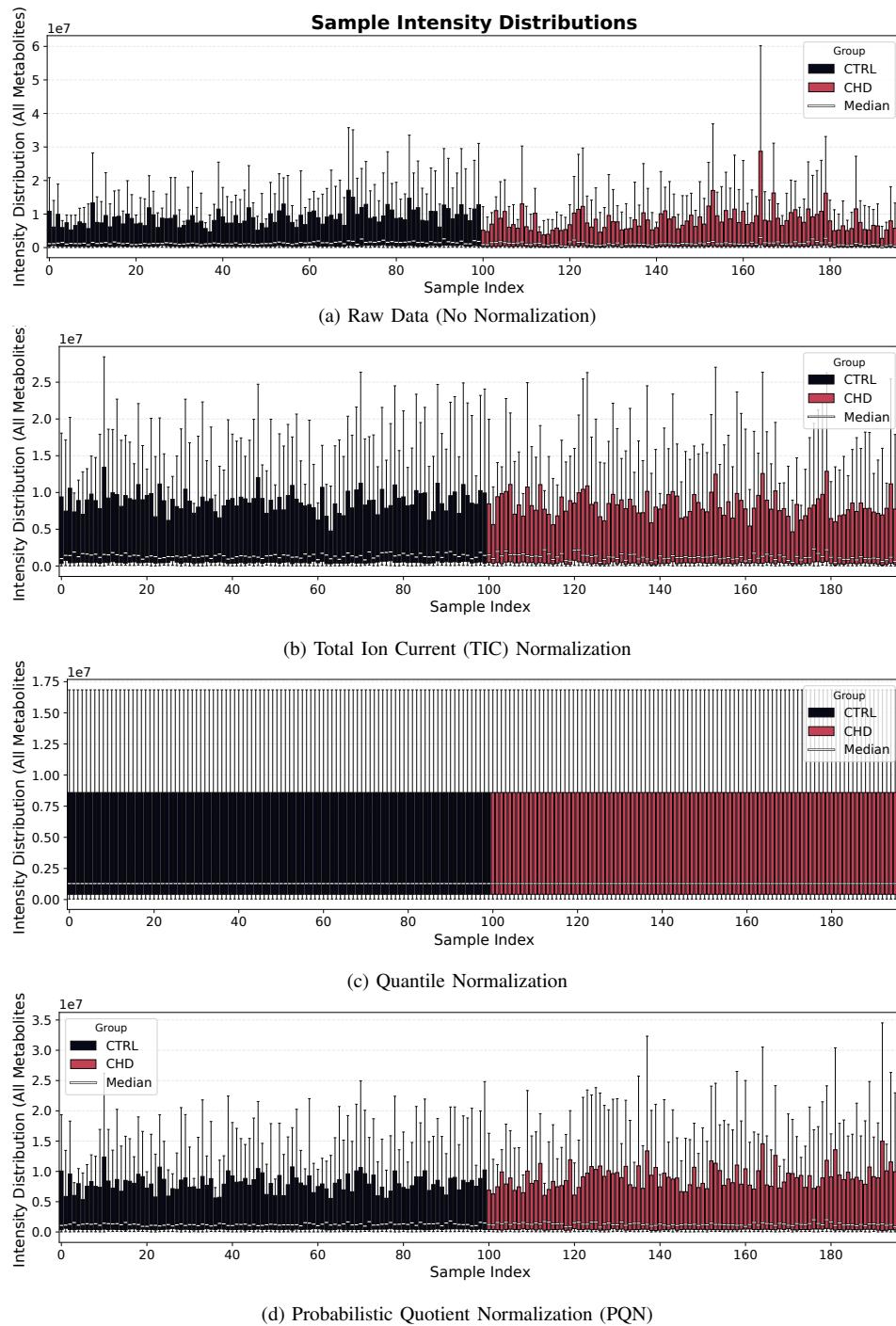


Fig. 3: **Comparison of Normalization Strategies for ESI- Dataset.** Boxplots representing the global intensity distribution of all samples. (a) Raw data showing significant systematic variation (e.g., dilution effects). (b) TIC normalization, acting on the total sum. (c) Quantile normalization, forcing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

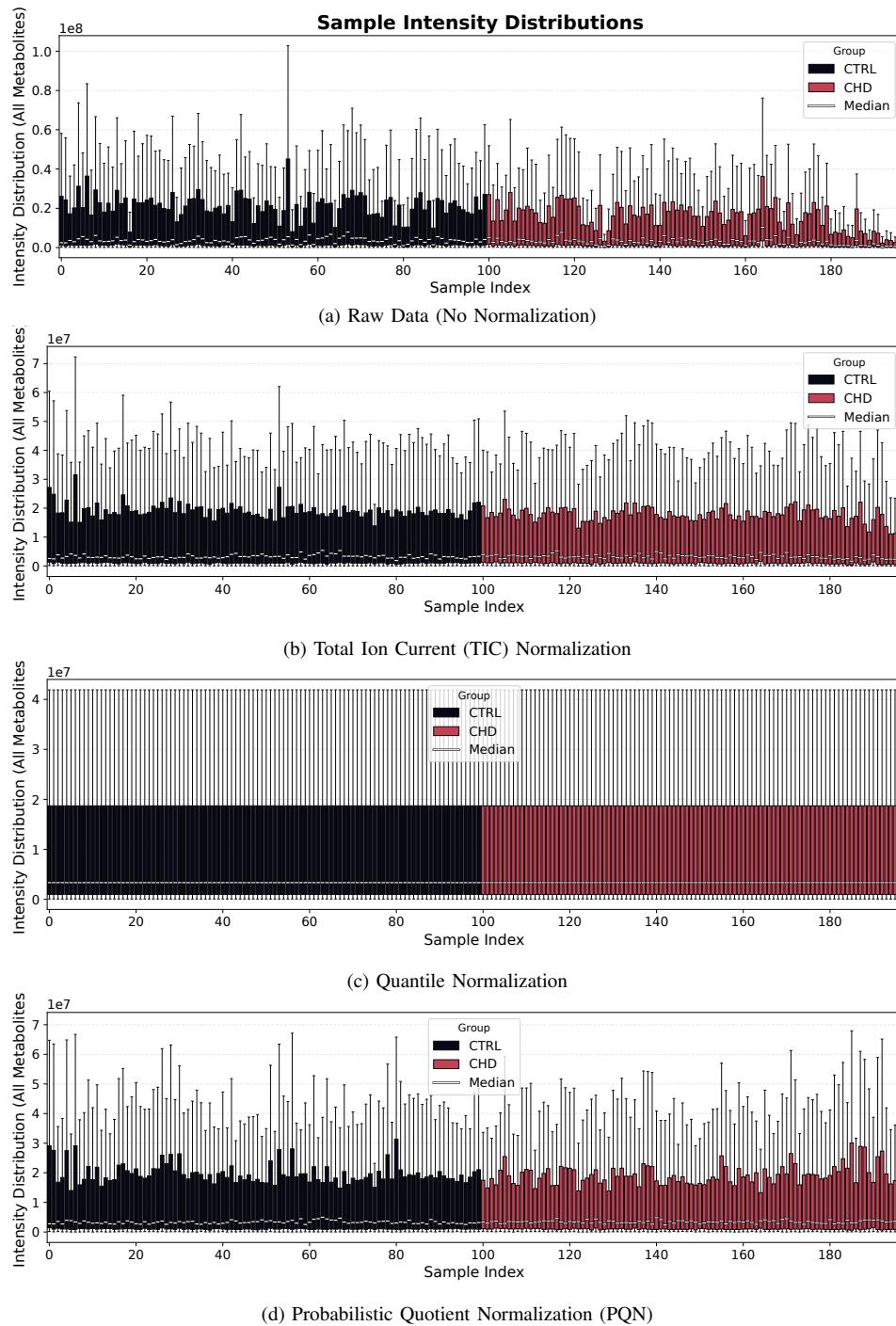


Fig. 4: **Comparison of Normalization Strategies for ESI+ Dataset.** (a) Raw data distribution. (b) TIC normalization results. (c) Quantile normalization results showing aggressive distribution alignment. (d) PQN results, selected as the optimal compromise for downstream analysis.

2.3.2 Data Transformation

Following the assessment of normalization strategies, the distribution of the intensity values was evaluated to satisfy the assumptions of normality required by multivariate statistical methods (e.g., PCA, PLS-DA) and parametric univariate tests (e.g., t-test). LC-MS metabolomics data typically exhibit a right-skewed distribution, where the majority of signals have low intensity, while a few highly abundant metabolites stretch the dynamic range, potentially dominating the variance.

We evaluated different variance-stabilizing transformations, including power transformations (Square Root, Cube Root) and logarithmic transformations (Log2, Natural Logarithm, Log10). Visual inspection of the global density plots revealed that power transformations were insufficient to correct the skewness of the distributions (data not shown). Conversely, all logarithmic transformations effectively compressed the high-intensity values and expanded the low-intensity range, resulting in a distribution

approximating a Gaussian curve.

Since Log2, Ln, and Log10 produced equivalent distributional shapes differing only in scale, **Log10 transformation** was selected as the standard method for this study. This transformation is widely accepted in mass spectrometry as it renders orders of magnitude easily interpretable while effectively symmetrizing the data distribution.

Figure 5 and Figure 6 illustrate the comparative analysis between the non-transformed data and the Log10-transformed data for the ESI- and ESI+ datasets, respectively. In the transformed data, the empirical density (solid line) shows a significant overlap with the theoretical Gaussian distribution (dashed line), confirming the efficacy of the transformation.

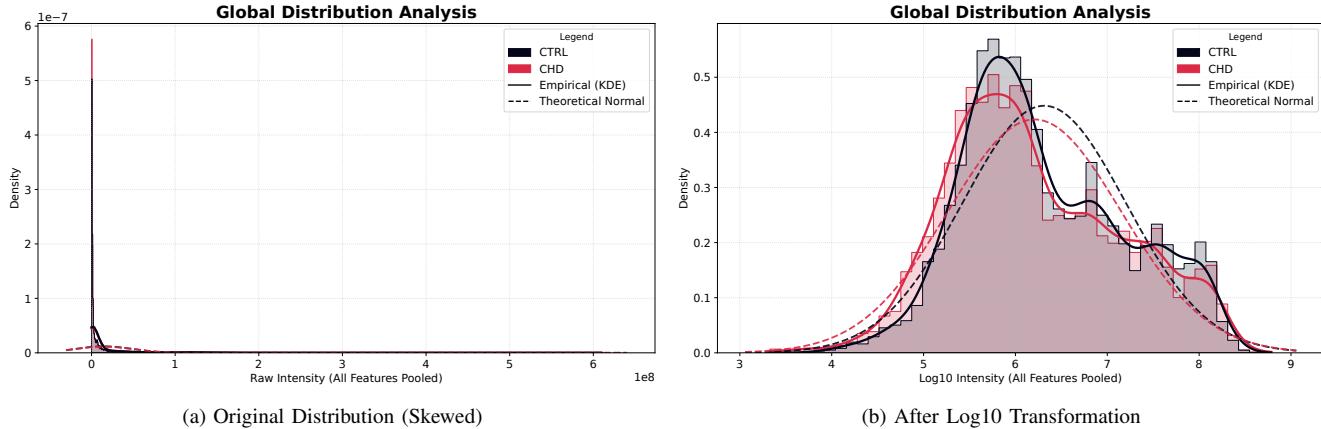


Fig. 5: Effect of Log10 Transformation on ESI- Dataset. Global density plots pooling all feature intensities. (a) The original data distribution is highly right-skewed, deviating significantly from the theoretical normal distribution (dashed line). (b) Log10 transformation successfully centers the distribution, achieving a Gaussian-like shape suitable for multivariate analysis.

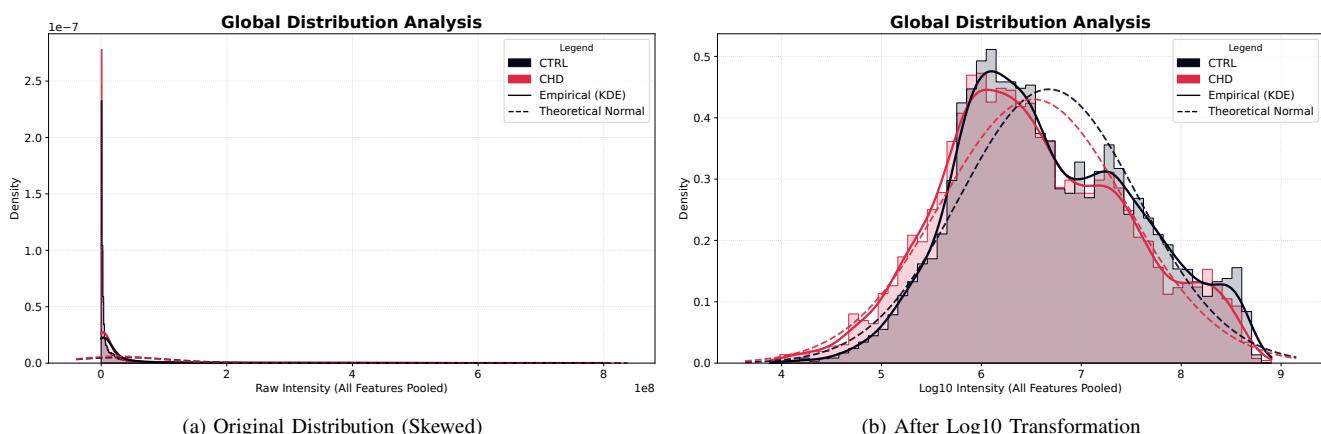


Fig. 6: Effect of Log10 Transformation on ESI+ Dataset. (a) The raw positive mode data exhibits a pronounced right-skewness. (b) Log10 transformation corrects the skewness, aligning the empirical density (solid line) with the theoretical Gaussian curve (dashed line).

2.3.3 Data Scaling

The final step of the pre-processing pipeline involved data scaling. In metabolomics, metabolite intensities can span several orders of magnitude. Without proper scaling, variables with high abundance and large variance would naturally dominate multivariate models based on variance maximization (e.g., PCA), biasing the results and potentially masking significant biological variations present in low-abundance metabolites.

To address this issue, **Autoscaling** (Unit Variance Scaling) was applied to the normalized and transformed data. This method involves mean-centering each variable and dividing it by its standard deviation. As a result, all metabolites are scaled to have a mean of zero and a standard deviation of one, ensuring that each feature contributes equally to the statistical model regardless of its absolute concentration.

Figure 7 and Figure 8 demonstrate the effect of autoscaling on the feature distributions for ESI- and ESI+ datasets, respectively. Before scaling (Panel a), the variables exhibit disparate ranges of intensity. After autoscaling (Panel b), all features are comparable, centered around zero with standardized variance, making the dataset suitable for unsupervised and supervised modeling.

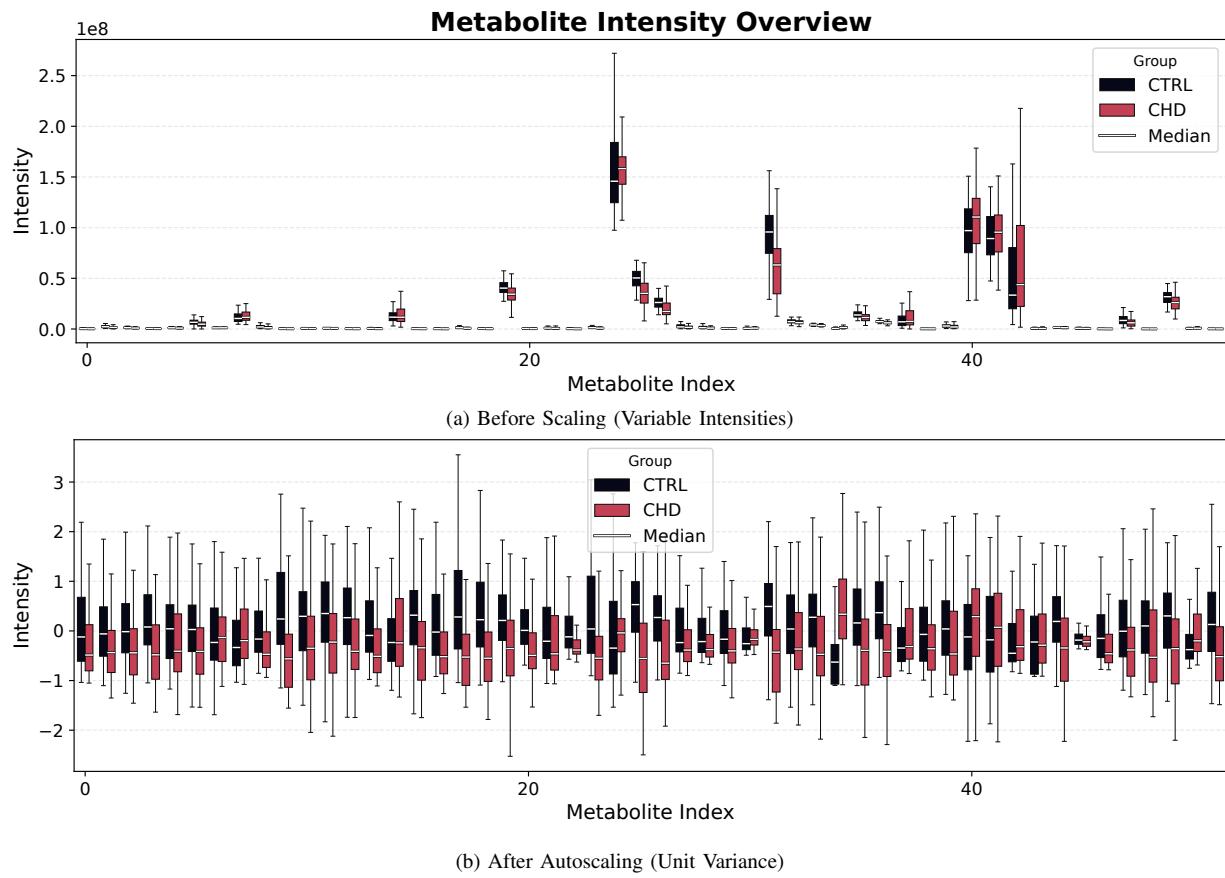


Fig. 7: Effect of Autoscaling on ESI- Features. Boxplots representing the distribution of individual metabolites (features). (a) Original features show highly variable ranges and variances. (b) After autoscaling, all features are mean-centered with unit variance, ensuring equal weight in multivariate analysis.

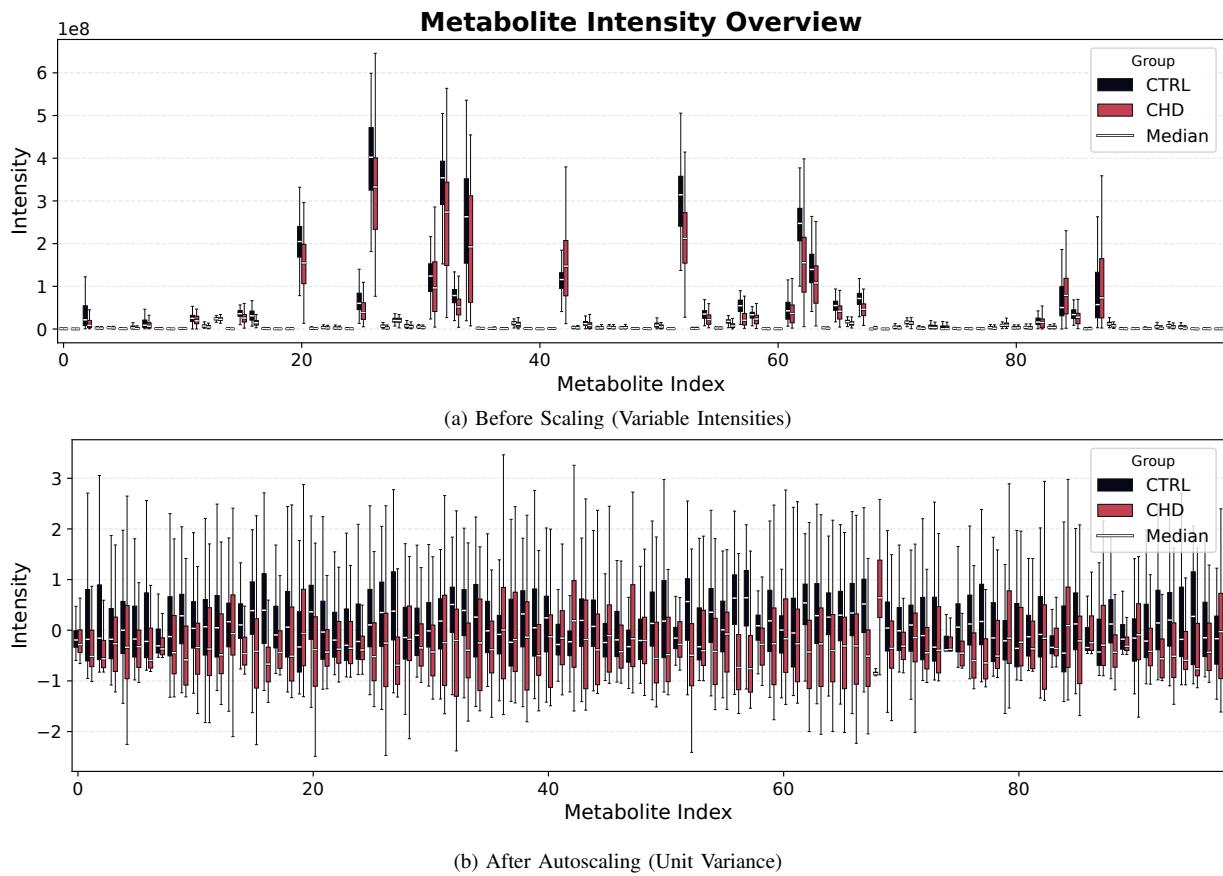


Fig. 8: **Effect of Autoscaling on ESI+ Features.** (a) The disparate scales of metabolite intensities in the positive mode. (b) The homogenized feature space achieved through autoscaling.

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 Data Fusion strategies

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

2.7 Statistical analysis and 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 Technology stack

2.8.1 NumPy

An essential Python package for scientific computing that provides support for large and multi-dimensional arrays and matrices. It serves as a backbone for the linear algebra calculations and computations, such as Frobenius norms for data fusion and the generation of coordinate matrices for visualization grids.

2.8.2 Pandas

Library for data analysis and manipulation, offering a suitable solution of DataFrames. It was broadly utilized to load and handle raw CSV files in a more suited version to the various design phases, including analysis and processing.

2.8.3 *Scikit-learn*

A comprehensive machine learning library broadly used for the data analysis and modelling. In this project it was used as a central piece of the data processing and preparation, including dimensionality reduction via PCA, scaling, outlier detection algorithms, and training predictive models, along with evaluating their performance.

2.8.4 *Statsmodels*

Python module providing statistical models functionalities, useful for conducting statistical tests. It was used in the univariate analysis, to guarantee a statistical validity of the identified biomarkers, by correcting for multiple hypothesis testing errors using the Benjamini-Hochberg False Discovery Rate (FDR) procedure.

2.8.5 *SciPy*

It is a scientific computing ecosystem library, adopted for statistical and probability functions. It provides the underlying algorithms for performing univariate hypothesis tests, such as Student's t-test, fitting theoretical normal distributions to data and calculating critical F-distribution values to establish decision boundaries in classification models.

2.8.6 *Matplotlib*

A complete Python library used for the creation of tailored and faultless publication-quality visualizations. It was extensively adopted in this project to generate the graphics in order to visually estimate the quality and performance of all the tested algorithms, while allowing for fine-grained control over aesthetic elements.

2.8.7 *Seaborn*

A statistical data visualization library based on Matplotlib, mainly used for the creation of more informative and specific graphs. It was used to generate clustered and regression plots that, equivalently to Matplotlib, automatically handling color mapping and legend generation to effectively communicate distribution patterns and class separations.

2.8.8 *tqdm*

A Python library used for creating a visual hint and monitoring the process with a practical loading bar.

2.8.9 *os*

A standard library module that provides a portable way of using operating system-dependent functionality, essential in this project to handle all the file management system operations.

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).

4.1 Future work

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.