# Evaluating Diagnostic Signatures of Crohn's Disease Across Biofluid Samples Using Machine Learning


## Abstract

This study evaluated breath, blood, faeces, and urine samples for their diagnostic potential in Crohn's disease using clustering and classification methods. Faeces consistently outperformed other sample types, exhibiting strong separability and classification accuracy. However, the small sample size relative to the high dimensionality of the data introduces uncertainty, highlighting the need for further validation with larger datasets.


## Introduction

The aim of this work is to determine which biological sample type—breath, blood, faeces, or urine—offers the best diagnostic power for identifying Crohn's Disease (CD). This analysis seeks to provide insights into the diagnostic potential of these samples, leveraging machine learning models to evaluate and compare their performance. The data used for this study was prepared by Dr. Michael Cauchi and consists of metabolite profiles derived from gas chromatography-mass spectrometry (GC-MS) analysis.

The dataset includes thousands of metabolite features for each sample, split into two groups: CD patients and healthy controls (CTRL). The dataset of comparison between CD and CTRL (CDvCTRL) was chosen for the analysis for simplicity. Other comparisons, such as CDvALL, introduce additional complexity due to mixed-class definitions that could dilute the diagnostic signal. Starting with these well-defined groups allowed for a clearer evaluation of the diagnostic potential of each sample type while ensuring robust and interpretable results.

## Methods

### *Data Loading and Preprocessing*

The initial stage of the analysis involved loading inspecting and preparing the GC-MS data for breath, blood, urine, and faecal samples. The raw data set used was initially prepared and provided by Dr. M Cauchi. The data set contained metabolite profiles and was in a .mat format, which required conversion into a more accessible and analyzable format. Implemented was a custom Python function provided by QMUL, designed to handle the specific structure of the .mat files. The gcparser function transformed the raw MATLAB data into pandas DataFrames by aligning the TIC matrix (XTIC) with samples as columns and retention times (RT) as rows. Sample names (SAM) and diagnostic class labels (CLASS) were flattened into lists for compatibility and ease of analysis, facilitating subsequent machine learning tasks

To streamline the workflow, the parsed DataFrames was organized into a hierarchical dictionary structure, grouped first by sample type (breath, blood, urine, or faeces) and then by comparison category (e.g., CD vs. CTRL). This organizational approach facilitated systematic access to the data for subsequent analyses.

### *Data Quality Assessment*

To evaluate the integrity of the data, a systematic quality check was implemented for each dataset. Using a custom function, missing values were assessed and outliers identified for all sample types. Missing values were quantified by summing the null entries in each dataset. Outliers were detected by calculating z-scores for all data points, with a threshold of $|z| > 3$ used to classify values as extreme. This process was repeated across all sample types and comparison categories (e.g., CD vs. CTRL), ensuring that the datasets were complete, and any potential anomalies were identified prior to further analysis.

### *Descriptive Statistics*

Descriptive statistics were computed to summarize the key properties of the data for each sample type and comparison category. For both CD and CTRL groups, the mean, median, standard deviation (std), and coefficient of variation (CV) were calculated for all retention times. These metrics were organized into new DataFrames for each sample type, capturing trends and variability within the data. Additionally, summary statistics, including count, range, and quartiles, were generated to provide an overall snapshot of the datasets. This analysis identified group-level differences and variability, forming the basis for interpreting the data's structure and variability prior to applying machine learning models.

### *Class Distribution*

To evaluate the balance between CD and CTRL classes in the datasets, the number of samples for each group across all sample types and comparisons was quantified. This involved iterating through the data dictionary to tally occurrences of "CTRL" and "CD" in the column names. The resulting counts were stored and visualized to provide an overview of the class distribution.

Bar plots were generated for each sample type, with proportions annotated on the plots to highlight any class imbalances. These visualizations provided a clear understanding of the dataset composition, identifying potential biases that could influence the performance of the machine learning models.

### *Principal Component Analysis (PCA)*

PCA was performed to reduce data dimensionality and uncover relationships between samples while preserving most of the variance. For each dataset, I first standardized the data to ensure all features had comparable scales, which is critical for PCA. Using the standardized data, I

applied PCA to extract the first two principal components, representing the majority of the variance in the data.

The results of the PCA were visualized as scatter plots, with each sample represented as a point according to its class (CD or CTRL). These plots highlighted clustering patterns and potential separations between the disease states. The percentage of variance explained by each principal component was annotated on the axes, providing insight into how much of the data's structure was captured by these components.

### *Silhouette Scores Using PCA Data*

To assess the degree of separation between CD and CTRL classes, silhouette scores were calculated using the silhouette_score function from scikit-learn. This analysis utilized the PCA-transformed data along with the corresponding class labels. The calculation was repeated for each sample type, providing a quantitative measure of how well-defined the clusters were within the reduced-dimensional space.

### *Machine Learning Models: Support Vector Machines (SVM)*

To classify samples as either CD or CTRL, a supervised learning approach using Support Vector Machines (SVM) with a linear kernel was implemented. Each dataset was preprocessed by transposing the data to ensure samples were represented as rows and features as columns. Class labels were encoded as binary values, with 1 for CD and 0 for CTRL.

The data was split into training (70%) and testing (30%) subsets to assess the model's generalizability to unseen data. Standardization was applied to both subsets, ensuring all features were on the same scale. The linear SVM model was trained on the standardized training set, and predictions were made on the testing set.

Performance evaluation metrics included accuracy and confusion matrices. Accuracy quantified the proportion of correct predictions, while confusion matrices detailed the distribution of true positives, true negatives, false positives, and false negatives. These metrics were calculated for each sample type, providing insight into the ability of SVMs to distinguish between CD and CTRL samples across different biological datasets.

### *Machine Learning Models: Random Forests (RF)*

For Random Forest classification, a RandomForestClassifier was used to predict CD and CTRL classes. The preprocessing steps included data transposition, class label encoding, splitting the data into training and testing sets, and standardization. The classifier was trained on the standardized training data, using its default settings to construct an ensemble of decision trees. Predictions were made for the testing set, and the model's performance was evaluated using accuracy and confusion matrices. The results were documented for each sample type and comparison.

### *Training and Evaluating SVM and Random Forest Models for Crohn's Disease Diagnosis*

For both SVM and Random Forest models, performance in diagnosing Crohn's disease was evaluated across all sample types and comparisons. The datasets were standardized and split into training and testing sets, ensuring consistent preprocessing. The SVM model, using a linear kernel, was trained on the standardized training data, and predictions, along with probability scores, were generated for the testing set.

Evaluation metrics, including accuracy, precision, recall, F1 score, and AUC-ROC, were calculated to comprehensively assess the model's classification ability and its effectiveness in distinguishing between CD and CTRL groups. These metrics provided insights into both the overall predictive performance and the balance between sensitivity and specificity.

The same procedure was applied to Random Forest models, with probabilities derived from the ensemble predictions used to compute the evaluation metrics. Results for both models were documented for each dataset, illustrating the diagnostic capabilities and comparative effectiveness of the models across different biological sample types.

### *ROC*

ROC curves and AUC were computed to evaluate the classification performance of the SVM and Random Forest models. Predicted probabilities for the testing data were used to calculate True Positive Rate (TPR) and False Positive Rate (FPR) at various thresholds, enabling the construction of the ROC curve. The AUC was computed to quantify overall model performance.

### *Bootstrap Validation of SVM and Random Forest*

To evaluate the reliability and generalizability of the SVM and Random Forest models, a bootstrap validation was performed with 500 iterations for all sample types and comparisons. Bootstrap validation involved resampling the dataset with replacement to generate multiple training and testing splits. Each resampled dataset was split into 70% training and 30% testing sets.

For each iteration, the training and testing data was standardized before training both models. The SVM model with a linear kernel and the Random Forest classifier were trained on the resampled training sets, and predictions were made on the corresponding testing sets. Accuracy scores were computed for both models in every iteration.

The results of the bootstrap validation were stored as distributions of accuracy scores for each model, sample type, and comparison. Mean accuracy and standard deviation were calculated to summarize the models' performance across all iterations, providing insight into their stability and effectiveness.

Stratified bootstrap validation was subsequently employed to improve the evaluation process, particularly to address potential class imbalances in the data. Using StratifiedShuffleSplit, the

data was split into training and testing subsets while maintaining the original proportions of CD and CTRL samples. This ensured that both classes were consistently represented in each iteration, reducing bias and variability in the results. For each stratified split, the SVM model was trained on the training subset, with predictions and probability scores generated for the testing subset. A comprehensive set of performance metrics was computed for each iteration, including accuracy, precision, recall, F1 score, and AUC-ROC. These metrics provided a more detailed evaluation of the SVM's classification performance, capturing its ability to differentiate between CD and CTRL samples across a balanced representation. The stratified validation approach was implemented for all sample types and comparisons, with results aggregated to compute mean and standard deviation for each metric.

# Results

### *EDA*

The data quality assessment confirmed the completeness of the datasets, with no missing values detected across all sample types, eliminating the need for imputation. However, outliers were identified using a z-score threshold of $|z| > 3$, with their counts varying across sample types: 1,632 in breath samples, 1,386 in blood, 1,450 in faeces, and 1,055 in urine. These outliers likely represent the natural variability inherent in biological data, potentially reflecting meaningful biological signals or noise depending on their context. Despite their presence, the datasets were considered robust enough for further analysis.

Moving from data quality to class distributions, slight imbalances were observed between the CTRL and CD groups across the sample types. These imbalances were not addressed using synthetic oversampling techniques like SMOTE, as the imbalances were not severe. Instead, stratified sampling was applied which will be explained later. This approach maintained the biological integrity of the data while ensuring that the machine learning models received representative and balanced input. While urine samples exhibited a slightly larger imbalance compared to other sample types, this was deemed manageable without the need for artificial data generation, as evidenced by the clear visualizations of the class distributions (Figure 1).

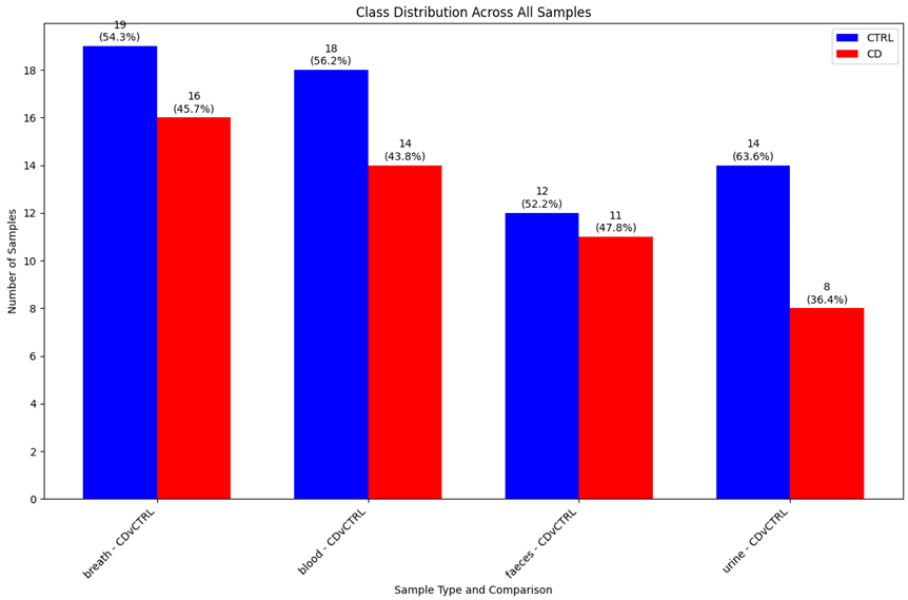


***Figure 1.*** *Class distribution across all sample types. Bar plots showing the number and proportion of Crohn's disease (CD) and control (CTRL) samples for each biological matrix (breath, blood, faeces, urine) under the CD vs. CTRL comparison. While all sample types exhibit a slight excess of CTRL samples, none display severe class imbalance. Percentages above each bar indicate the relative fraction of each class.*

Exploration through Principal Component Analysis (PCA) revealed varying degrees of clustering between the CD and CTRL groups across different sample types (Figure 2). Breath samples demonstrated the strongest separation, with the first two principal components explaining 93.62% of the variance (PC1: 82.27%, PC2: 11.35%), providing a clear distinction between the two groups. Faeces samples displayed moderate separation, with PC1 and PC2 accounting for 56.74% of the variance (PC1: 35.06%, PC2: 21.68%). Blood samples showed less pronounced separation, with PC1 and PC2 capturing 75.40% of the variance (PC1: 60.38%, PC2: 15.02%) and notable overlap between groups. Urine samples demonstrated the least separation, with PC1 and PC2 explaining 85.93% of the variance (PC1: 75.39%, PC2: 10.54%), and considerable overlap between the CD and CTRL groups.

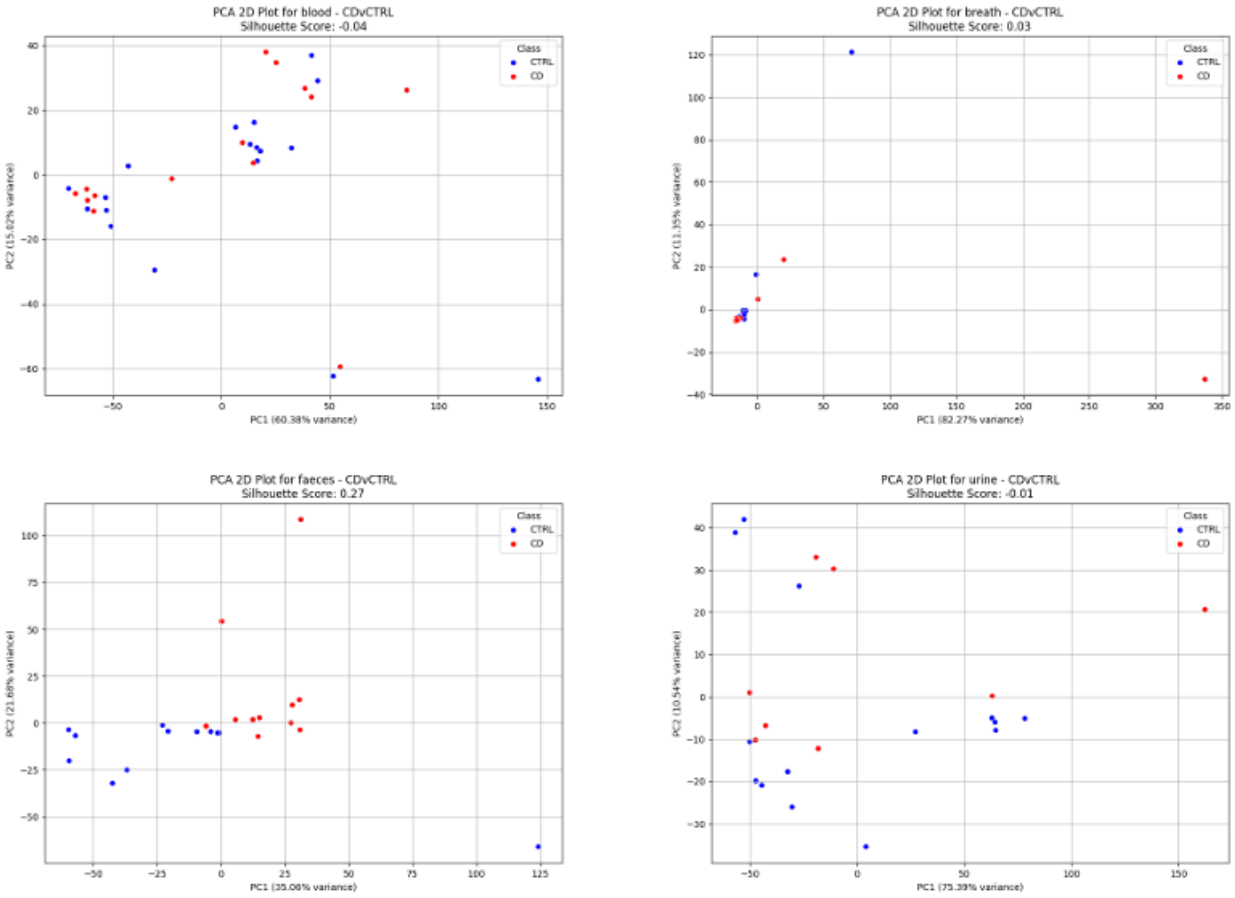


***Figure 2.*** *PCA-based visualization of sample separations and silhouette scores. Two-dimensional principal component analysis (PCA) scatter plots for each sample type (blood, breath, faeces, urine) color-coded by class (CD = red, CTRL = blue). The percentage of variance explained by PC1 and PC2 is noted on the axes. Silhouette scores, displayed in each panel, quantify cluster separation quality. Faeces show the highest silhouette score (0.27), indicating better-defined class distinctions, while blood, breath, and urine display lower or near-zero scores, suggesting poorer separation between CD and CTRL groups.*

To further evaluate the clustering quality, silhouette scores were calculated based on the PCA-transformed data (Figure 2). These scores quantify the cohesion within clusters and their separation from other clusters, ranging from -1 (poor separation) to +1 (well-defined clusters). The results somewhat supported the insights from PCA. Faeces samples exhibited the highest silhouette score (0.268), indicating well-defined clusters and minimal overlap between CD and CTRL groups. In contrast, breath samples had a low positive score (0.034), suggesting only marginal clustering. Blood and urine samples had negative silhouette scores (-0.038 and -0.014, respectively), reflecting overlapping clusters and poorly defined group boundaries. These findings highlight faeces as a promising candidate for diagnostic purposes, while breath, blood, and urine may require additional exploration or combination with other datasets to provide meaningful diagnostic insights.

### *Machine Learning Classification (SVM and Random Forest)*

The performance of Support Vector Machines (SVM) and Random Forest (RF) classifiers was evaluated for their ability to distinguish between CD and CTRL groups across the four sample types. Accuracy and confusion matrices provide a comparative perspective on the predictive power of the two models in the context of diagnosing Crohn's disease.

Faeces samples consistently demonstrated the highest classification performance across both models. The Random Forest classifier achieved perfect accuracy (100%), correctly classifying all CD and CTRL samples with no misclassifications, as evidenced by the confusion matrices (Figure 4). SVM also performed well with faeces, achieving an accuracy of 85.7%, with only one misclassification. These results align with previous analyses, confirming faeces as the most diagnostically relevant sample type (Figure 3 for SVM, Figure 4 for RF).

Breath samples showed moderate performance, with both SVM and RF achieving the same accuracy of 63.6%. While the SVM model misclassified a slightly larger proportion of samples compared to RF, both models face challenges in distinguishing between the two groups (Figure 3 for SVM, Figure 4 for RF). Blood samples exhibited the weakest performance, with SVM and RF achieving accuracies of 50.0% and 40.0%, respectively, highlighting the limited diagnostic potential of blood as a standalone sample type.

Urine samples showed contrasting results between the two models. SVM achieved a modest accuracy of 57.1%, with multiple misclassifications, whereas RF outperformed SVM, achieving an accuracy of 85.7%. The RF confusion matrix reveals better-defined group separations for urine, suggesting that RF might be more effective in leveraging the features in urine data (Figure 3 for SVM, Figure 4 for RF).

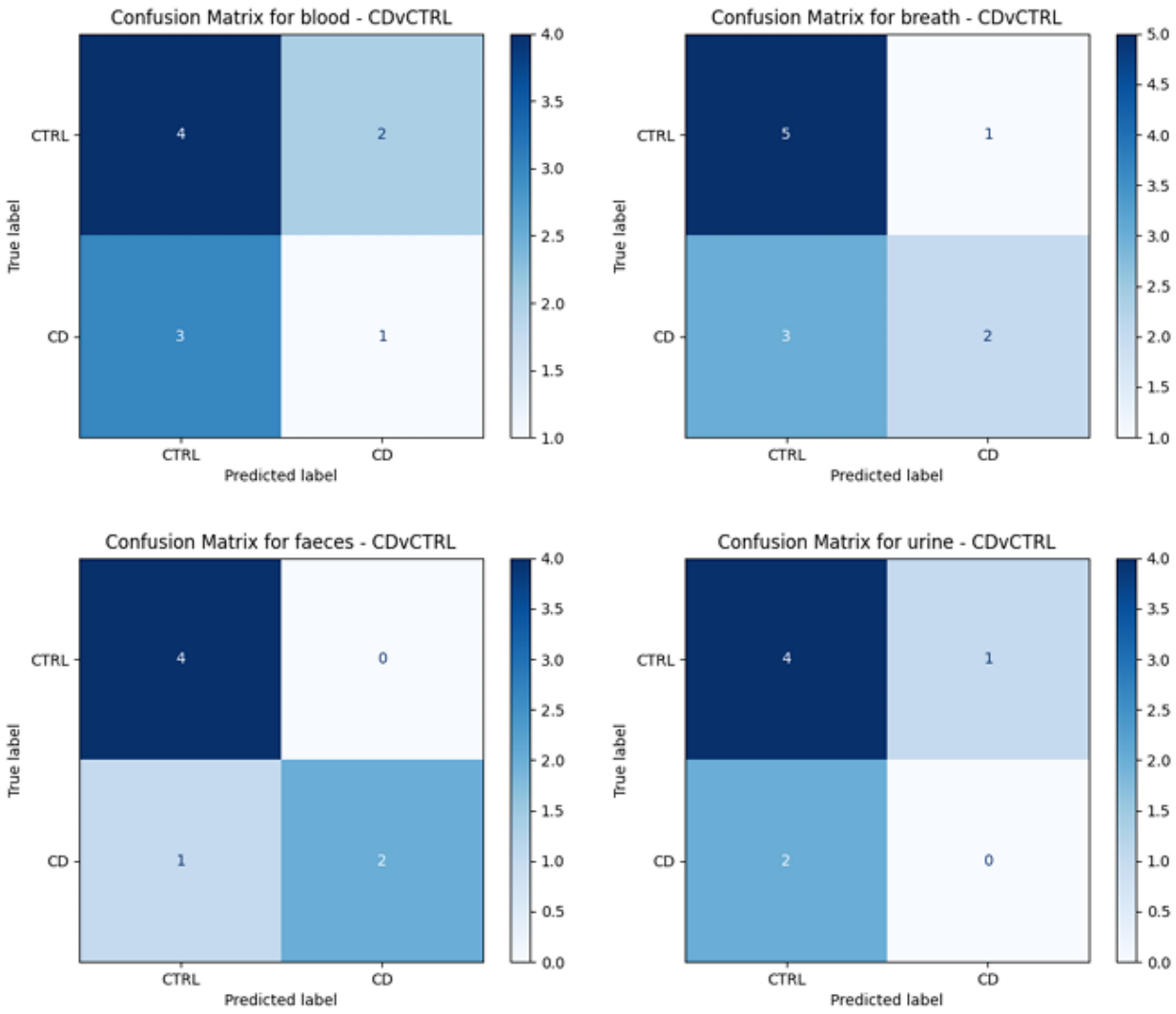


***Figure 3.*** *SVM confusion matrices for CD vs. CTRL classification across sample types. Each matrix compares predicted versus true classes for the Support Vector Machine (SVM) classifier applied to blood, breath, faeces, and urine datasets. Darker cells along the diagonal represent correct classifications. Faeces yield fewer misclassifications, supporting its stronger diagnostic potential, whereas blood, breath, and urine show more off-diagonal counts, reflecting lower discriminative power.*

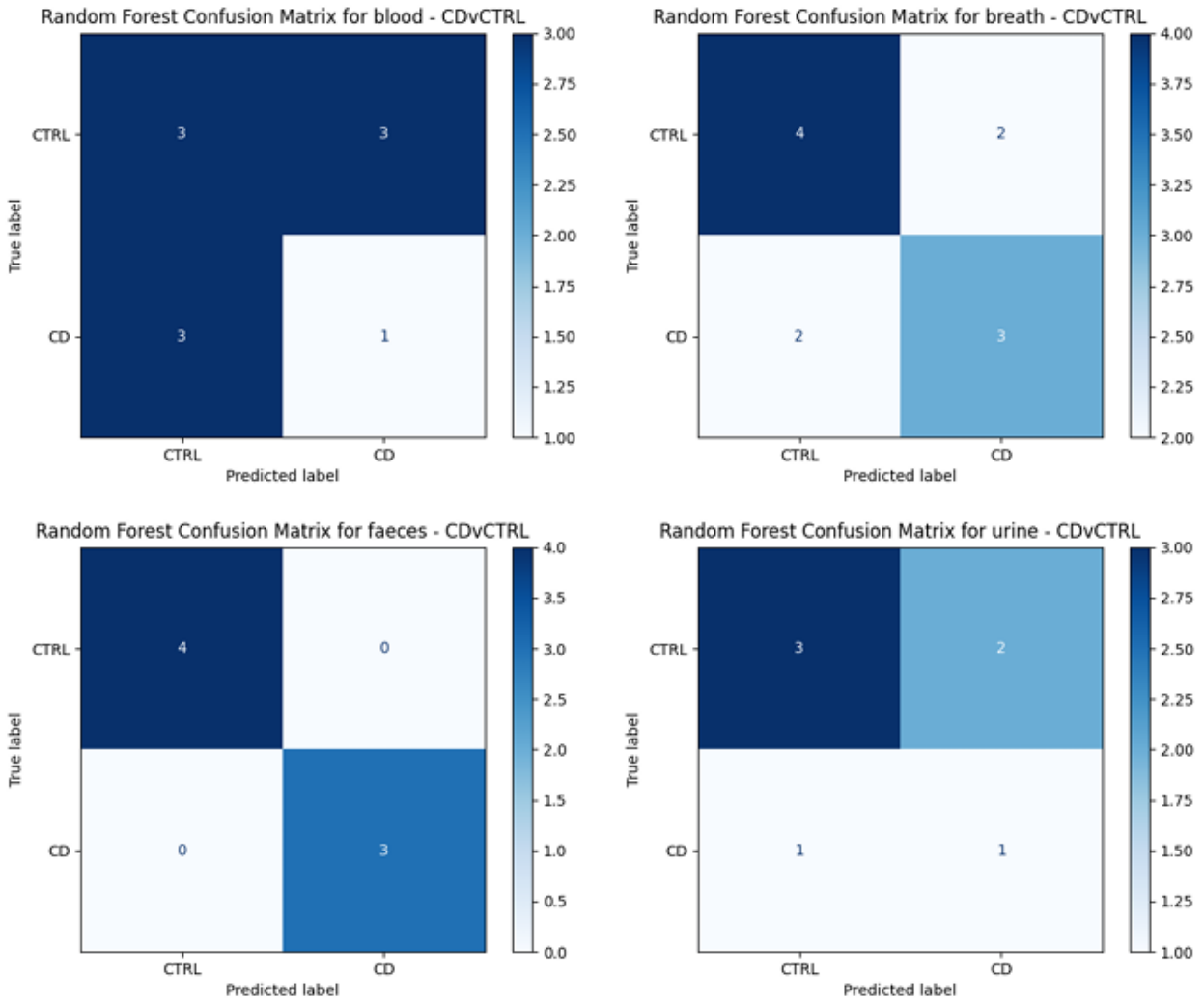


***Figure 4.*** *Random Forest confusion matrices for CD vs. CTRL classification across sample types. Confusion matrices depict the performance of a Random Forest classifier on blood, breath, faeces, and urine samples. Correct classifications cluster along the diagonal. Faeces samples demonstrate near-perfect classification, underscoring the robust disease signal in this matrix. By contrast, blood and urine show higher misclassification rates, and breath yields moderate performance.*

### *ROC*

The ROC curve analyses (figure 5) revealed that faeces, breath, and urine samples demonstrated diagnostic performance above random chance (50%) for at least one model, highlighting their potential utility. In contrast, blood samples showed limited diagnostic value, with one model performing at or below random chance. These findings underscore the variability in diagnostic potential across sample types, with faeces emerging as the most promising candidate.

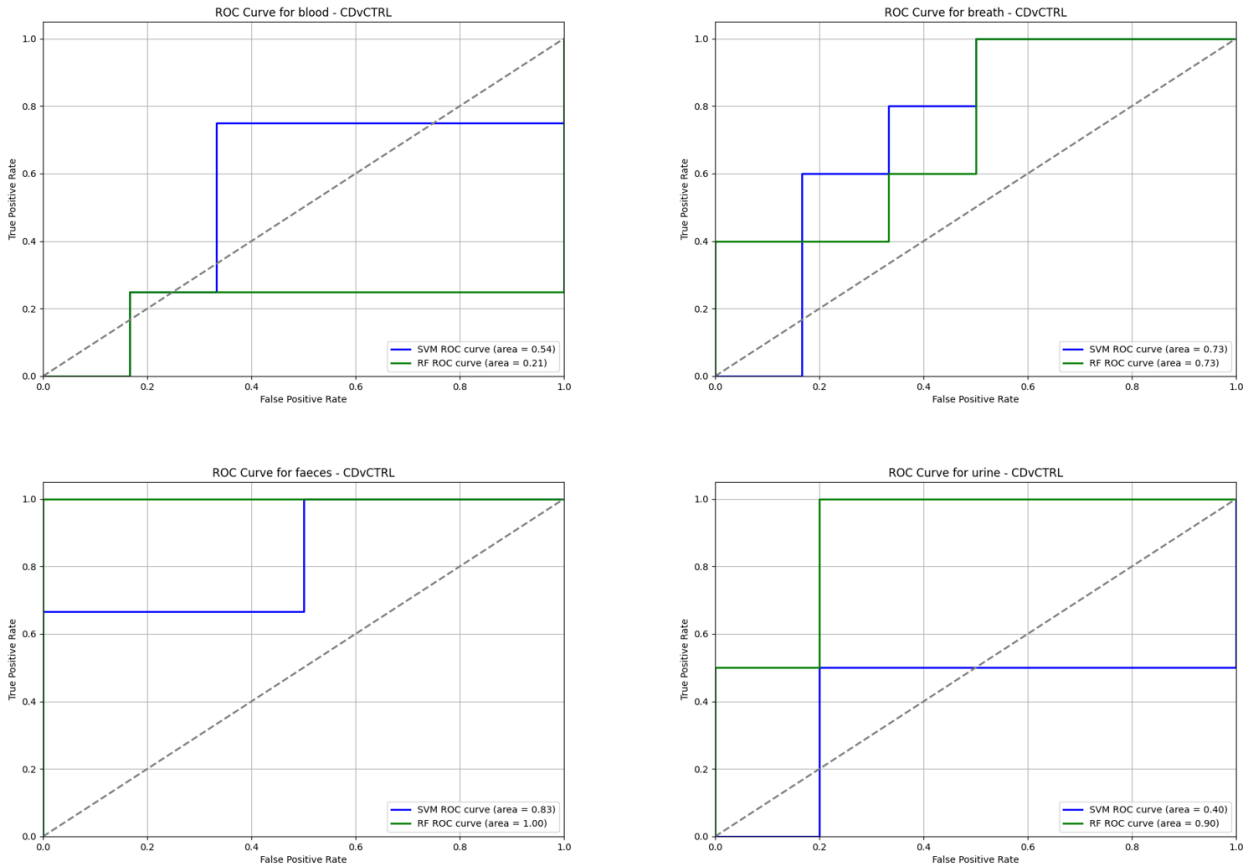


*Figure 5. Receiver Operating Characteristic (ROC) curves comparing Support Vector Machine (SVM, blue) and Random Forest (RF, green) classifiers for distinguishing Crohn's disease (CD) from controls (CTRL) in the sample types. The diagonal dashed line indicates the line of no-discrimination (AUC = 0.5). Each plot reports the Area Under the Curve (AUC) for both models, reflecting their diagnostic utility. Faeces (C) exhibit the highest AUC values, indicating superior discriminative performance, while blood (A) shows near-random classification accuracy. Breath (B) and urine (D) yield intermediate results, suggesting modest diagnostic value.*

### *Bootstrap Validation*

The significantly higher mean accuracy observed during bootstrap validation for both SVM and Random Forest models raised concerns regarding the generalizability of the models. The breath and faeces samples, for example, achieved mean accuracies of 79.4% and 88.3% for SVM, and 81.6% and 90.9% for Random Forest, respectively. While these results are promising, they stand in contrast to the lower accuracies observed during the initial evaluation, particularly for breath and urine samples, where clustering and silhouette scores indicated limited group separability.

The high accuracies in bootstrap validation can be attributed to its methodology, which involves resampling with replacement from the dataset. This approach increases the effective size of the training data, potentially allowing the models to learn better. However, the lack of explicit measures to address class imbalance may lead to optimistic results that do not reflect real-world diagnostic performance.

To mitigate these concerns and ensure more reliable performance evaluation, stratified bootstrap validation was employed in the subsequent analysis. This method preserves class ratios during resampling, ensuring more balanced and representative subsets for training and testing. The distributions of accuracy from the bootstrap validation for each sample type and model are depicted in histogram form, revealing how often models achieve particular accuracy ranges and highlighting differences in their stability and reliability across sample types (Figure 5).

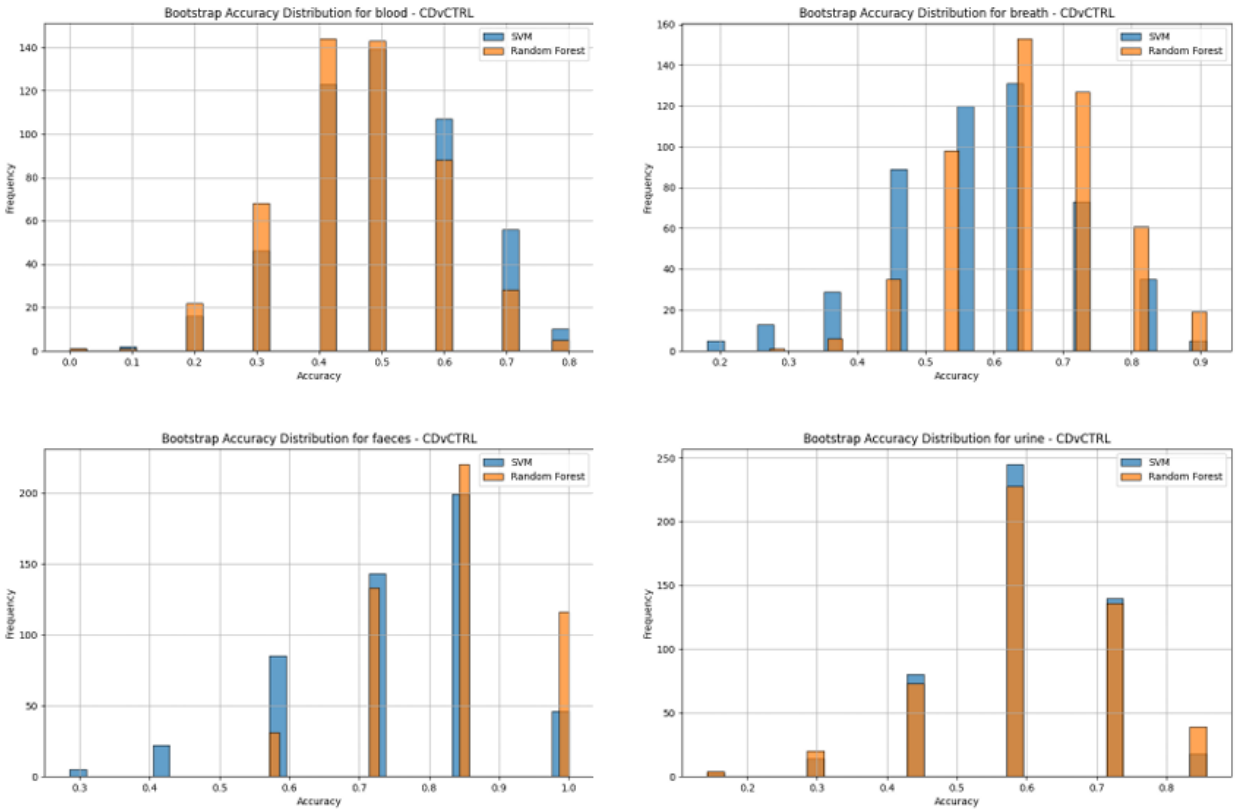


***Figure 6.*** *Bootstrap accuracy distributions for SVM and Random Forest models. Histograms illustrating the distribution of accuracy scores from 500 bootstrap validation iterations for blood, breath, faeces, and urine datasets. Blue bars represent SVM results, and orange bars represent Random Forest. Faeces display accuracy distributions shifted towards higher values, confirming more stable and reliable diagnostic performance. In contrast, blood, breath, and urine show broader distributions with lower median accuracies, indicating less consistent predictive ability and weaker class separation.*

Stratified bootstrap validation provided a more balanced evaluation of SVM and Random Forest models across the datasets for the sample types. For breath, SVM achieved a mean accuracy of 58.1% with a mean precision of 54.2%, recall of 62.8%, F1-score of 56.6%, and AUC-ROC of 57.9%. Random Forest performed slightly better, with a mean accuracy of 65.7%, mean precision of 66.9%, recall of 56.5%, F1-score of 58.6%, and AUC-ROC of 70.7%. These results suggest moderate separability of classes, with Random Forest displaying a slight edge due to its higher recall and AUC-ROC.

For blood, the performance of both models was limited. SVM yielded a mean accuracy of 49.5%, mean precision of 38.4%, recall of 43.1%, F1-score of 39.0%, and AUC-ROC of 45.7%. Random Forest showed similar results with a mean accuracy of 46.1%, mean precision of 34.0%, recall of 39.5%, F1-score of 35.2%, and AUC-ROC of 46.2%. These results indicate poor separability between CD and CTRL classes in blood samples, with metrics close to random chance.

For faeces, both models demonstrated the best performance among the sample types. SVM achieved a mean accuracy of 75.6%, mean precision of 73.1%, recall of 78.6%, F1-score of 73.0%, and AUC-ROC of 84.6%. Random Forest performed significantly better with a mean accuracy of 83.5%, mean precision of 79.3%, recall of 89.8%, F1-score of 82.7%, and AUC-ROC of 92.5%. The high recall and AUC-ROC for Random Forest suggest that faeces samples provide strong diagnostic potential for distinguishing between CD and CTRL classes.

For urine, the models showed limited effectiveness. SVM achieved a mean accuracy of 58.8%, mean precision of 42.5%, recall of 23.6%, F1-score of 28.7%, and AUC-ROC of 46.2%. Random Forest slightly outperformed SVM with a mean accuracy of 59.7%, mean precision of 42.5%, recall of 22.9%, F1-score of 28.1%, and AUC-ROC of 65.1%. Despite slightly higher AUC-ROC values for Random Forest, both models struggled with low recall and F1-scores, reflecting limited separability of the classes in urine samples.

***Table 1.*** *Comparative performance metrics of SVM and Random Forest models for diagnosing Crohn's disease across different biological sample types. Each cell reports mean values obtained from stratified bootstrap validation (500 iterations), including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Higher values indicate better model performance. Faeces emerge as the most diagnostic matrix, with Random Forests achieving especially high accuracy, recall, and AUC-ROC, underscoring the strong discriminatory signal present in faecal metabolite profiles. Other sample types, particularly blood and urine, show weaker classification metrics, suggesting lower diagnostic utility under the evaluated conditions.*

| Sample Type | Model | Mean Accuracy (%) | Mean Precision (%) | Mean Recall (%) | Mean F1-Score (%) | Mean AUC-ROC (%) |
|---|---|---|---|---|---|---|
| Breath | SVM | 58.1 | 54.2 | 62.8 | 56.6 | 57.9 |
| | Random Forest | 65.7 | 66.9 | 56.5 | 58.6 | 70.7 |
| Blood | SVM | 49.5 | 38.4 | 43.1 | 39 | 45.7 |
| | Random Forest | 46.1 | 34 | 39.5 | 35.2 | 46.2 |
| Faeces | SVM | 75.6 | 73.1 | 78.6 | 73 | 84.6 |
| | Random Forest | 83.5 | 79.3 | 89.8 | 82.7 | 92.5 |
| Urine | SVM | 58.8 | 42.5 | 23.6 | 28.7 | 46.2 |
| | Random Forest | 59.7 | 42.5 | 22.9 | 28.1 | 65.1 |

The stratified bootstrap approach addressed the issue of imbalanced class distributions seen in earlier validations. This led to more reliable metrics, especially for models applied to datasets with skewed class representation. Faeces emerged as the most diagnostically informative sample type, with Random Forest significantly outperforming SVM in its ability to distinguish between CD and CTRL classes. In contrast, blood samples showed the weakest performance, highlighting a lack of distinct class separability. Random Forest consistently demonstrated better recall and AUC-ROC values compared to SVM, particularly in breath and faeces samples, which underscores its ability to capture complex patterns in the data. Further contextualization of these results will be discussed in the next section by integrating clustering outcomes to explore the underlying separability of CD and CTRL classes.

# Discussion

The aim of this report was to determine which biological sample type provides the most diagnostic signature for Crohn's disease (CD). The results indicate that faecal samples emerge as the strongest candidate, with Random Forest achieving perfect accuracy and faeces exhibiting the highest silhouette score and clearest separability in PCA. To clarify, the separation in PCA plots reflects variance in the data but does not directly imply better diagnostic potential but perhaps that the faecal sample may have been a better candidate for such analysis. The literature says that faeces are likely to capture microbial and metabolic changes directly associated with CD pathology, making it a rich source of diagnostic markers. Vich Vila et al. (2023)

In contrast, blood demonstrated the weakest diagnostic potential, reflected in lower accuracy and clustering metrics. Breath and urine showed intermediate performance, suggesting some diagnostic utility, though less pronounced than faeces. These findings suggest a gradient of diagnostic informativeness, with faeces being the most robust, followed by urine and breath, and finally blood.

Despite these findings, several factors introduce uncertainty into the analysis. The small sample size, with fewer than 20 samples per class for each biological sample, may have skewed the performance of machine learning methods. While the datasets contain thousands of features, the limited number of samples increases the risk of overfitting and reduces the generalizability of the models. This is particularly critical for high-dimensional data, where the risk of identifying spurious patterns increases with fewer samples.

Moreover, the imbalance in sample distribution across classes could have influenced clustering and classification results, even with techniques like stratified bootstrap validation. These uncertainties highlight the need for larger, more balanced datasets to confirm the observed trends and reduce the influence of noise or artifacts in the data.

Finally, the choice to utilize CDvCTRL for simplicity instead of or in addition to CD vs other Inflammatory bowel diseases may or may not have improved the outcomes, which underscores the need for further experiments to validate and expand on such implementations.

### Conclusion

In conclusion, faeces demonstrate the strongest diagnostic potential for Crohn's disease among the four biological samples analyzed, showing consistent clustering, high classification accuracy, and robust separability. However, the small sample size and high dimensionality of the data introduce significant uncertainty, necessitating further validation with larger, more balanced datasets to ensure the reliability and generalizability of these findings.

## Reference

Vich Vila, A. *et al.* (2023) 'Faecal metabolome and its determinants in inflammatory bowel disease', *Gut*, 72(8), pp. 1472–1485. Available at: https://doi.org/10.1136/gutjnl-2022-328048.