Comparative Analysis of Biological Samples for Crohn's Disease Diagnosis Using Machine Learning Techniques

Abstract

The goal of this study was to use machine learning to determine which biological sample type—breath, blood, urine, or faeces—provided the most accurate diagnostic signal for Crohn's disease. Gas chromatography-mass spectrometry (GC-MS) data from each sample type were analysed using Support Vector Machine (SVM) and Random Forest models to determine whether the samples were CD or control. The models' robustness was tested using Bootstrap validation and permutation tests. The study found that stool samples included the most diversified biochemical fingerprints, with the SVM model showing high accuracy and consistency, especially when separating CD patients from healthy controls. Urine samples showed promise, while blood and breath samples were less successful due to poorer diagnostic accuracy and increased susceptibility to overfitting. Finally, faecal metabolomics showed the most potential as a non-invasive diagnostic technique for CD, exceeding other sample types and producing significantly better findings than chance. These findings support the use of faecal samples in the clinical diagnosis and surveillance of Crohn's disease, with future developments potential through the application of additional omics techniques.

Roxana Andreea Bosnea

BIO720P - AI AND DATA SCIENCE IN BIOLOGY

CONTENTS

INTRODUCTION

Crohn's disease (CD) is a chronic inflammatory disorder of the gastrointestinal tract that poses major diagnostic difficulties due to its complicated and diverse clinical symptoms. To make a clear diagnosis of CD, intrusive procedures such as endoscopy and biopsy are frequently required, which are not only uncomfortable for patients but also resource-intensive (Baumgart & Sandborn, 2012). Recent breakthroughs in metabolomics, notably the use of gas chromatography-mass spectrometry (GC-MS), provide a promising non-invasive method for studying metabolic processes in biological materials.

Finding out which biological sample – blood, breath, urine, or faeces – offers the most useful signature for the diagnosis of Crohn's disease is the aim of this study. These samples' GC-MS chromatograms will be used to categorise the various illness stages. This will be accomplished by using two machine learning techniques: *Random Forests* and *Support Vector Machines* (SVM). The GC-MS acquired metabolomic data, which may also function as CD biomarkers, illustrates the intricate interactions among metabolites across the body (Lloyd-Price, et al., 2019). After GC-MS data from four different but related types are prepared and imported, machine learning models will be used to assess each sample type's diagnostic potential. Permutation testing and bootstrap validation will also be used to make sure the models are reliable and that their performance is not the result of random variation.

METHODS

DATA ACQUISITION AND PREPROCESSING

Gas chromatography-mass spectrometry (GC-MS) analyses of four distinct biological sample types – blood, breath, faeces, and urine – provided the metabolomic data for the study. These samples were chosen with the intention of evaluating their capacity to provide Corhn's disease (CD) diagnostic signals. The **loadmat()** function of **scipy.io** package was used to import the GC-MS data, which was initially in the **.MAT** format used by MATLAB. This program imports the binary files into Python as dictionaries. The total ion counts (TIC) were determined for every sample throughout a variety of retention periods and were combined with metadata, such as sample names and class designations (CD patient or healthy control), in the GC-MS data.

For further analysis, the imported data was wrangled into Pandas DataFrame using a custom parsing tool called **gcparser**. It collects the TIC matrix, sample names, retention durations, and class labels from the MATLAB structures and following that, each sample's class is recognised in a different column, and the data is organised into a structure format with samples arranged as rows and retention lengths as columns. This preparation ensures the data is ready for the development of machine learning models.

CHROMATOGRAM VISUALISATION

The sample chromatograms were visualised before machine learning models were used in order to do proceed with exploratory data analysis. Chromatograms, which display the variation in ion counts as a function of retention time, provide a visual representation of the metabolomic profiles of the samples. The **plot_chromatograms** function was used to plot specific chromatograms from each dataset, enabling qualitative comparisons of CD samples with healthy controls ahead of time. The visualisation step is crucial in identifying any differences in the metabolic profiles between the classes.

MACHINE LEARNING MODEL DEVELOPMENT

The two machine learning models that were used to classify the samples were *Support Vector Machines (SVM)* and *Random Forests*. These models were chosen because of their ability to manage complex, high-dimensional data, such as that obtained from GC-MS analyses.

- **Support Vector Machine (SVM)**: The SVM model was implemented using the **SVC()** function of the **scikit-learn** library. To train the model and produce a hyperplane that maximises the margin between the two classes (CD vs control), GC-MS data were utilised. The SVM model was trained on the dataset using the **train_evaluate_svm** function, which also produced a confusion matrix and calculated the prediction accuracy on the same dataset.
- **Random Forest**: The **RandomForestClassifier()** function from **scikit-learn** was used to create the Random Forest model, a popular ensemble learning method. During training, this model builds a number of decision trees, and it outputs the classification – or class mode – from each individual tree. The model was trained on the dataset using the **train_evaluate_rf** function where it evaluated its accuracy and displayed the corresponding confusion matrix.

MODEL VALIDATION: BOOTSTRAP AND PERMUTATION TESTING

In order to ensure reliability of the classification models, there were two validation techniques that were employed: *bootstrap validation* and *permutation testing*.

- **Bootstrap Validation**: Bootstrap validation was done to estimate the model's performance stability by resampling the dataset with replacement. For each iteration, the dataset was randomly split into training and testing sets using the **train_test_split()** function, which by default allocates 70% of the data for training and 30% for testing. This process was then repeated 100 times, and the average accuracy was computed to provide the estimate of model performance. The **boostrap_validation** function facilitated in this process, and the distribution of accuracy was visualised using histograms.
- **Permutation Testing**: Permutation testing was done to see if the reported accuracy was higher than random chance. Using this method, the model was retrained and assessed using the permuted data after the class labels were randomly mixed. In addition, this process was carried out 100 times. The statistical significance of the model's performance was then ascertained by contrasting the accuracy distribution that resulted from the permuted data with the original accuracy distribution. Plotting the data as overlapping histograms was done using the **permutation_testing** function.

COMPARISON ACROSS SAMPLE TYPES

To find out which biological sample had the highest diagnostic value for Crohn's disease, the trained models were applied to GC-MS data from all four sample types (blood, breath, faeces, and urine). The accuracy of the Random Forest and SVM models was then validated using both bootstrap and permutation validation methods. Plots were generated for each type of sample to illustrate the model's performance over the various datasets. The information was gathered into a DataFrame and visualised to determine the optimal sample type for diagnosing Crohn's disease.

RESULTS & DISCUSSION

OVERVIEW OF SAMPLE TYPES AND PERFORMANCE METRICS

In this study, four different types of samples—blood, breath, faeces, and urine—were examined to see which produced the greatest diagnostic signature for Crohn's Disease (CD). The performance of each sample type was evaluated using two machine learning models: *Support Vector Machine* (SVM) and *Random Forest*. The accuracy of these models was tested in two groups: CD_All (all CD samples) and CD_Ctrl (CD samples vs control samples). Additionally, bootstrap and permutation tests were used to assess the robustness of the results.

BLOOD SAMPLES

Chromatogram Analysis

Chromatograms of the blood samples were generated using both the CD_All and CD_Ctrl databases. The chromatograms demonstrate chemical profiles taken from blood samples of Crohn's disease (CD) patients and controls. Figure 1 shows chromatograms for blood samples from the CD_All dataset. The retention lengths and ion counts provide information about the metabolites found in blood samples. Notable peaks, such as those at retention times of 7.67 minutes and 14.34 minutes, can be seen in both CD and control samples, albeit the intensity and existence of specific peaks differs between the two groups.

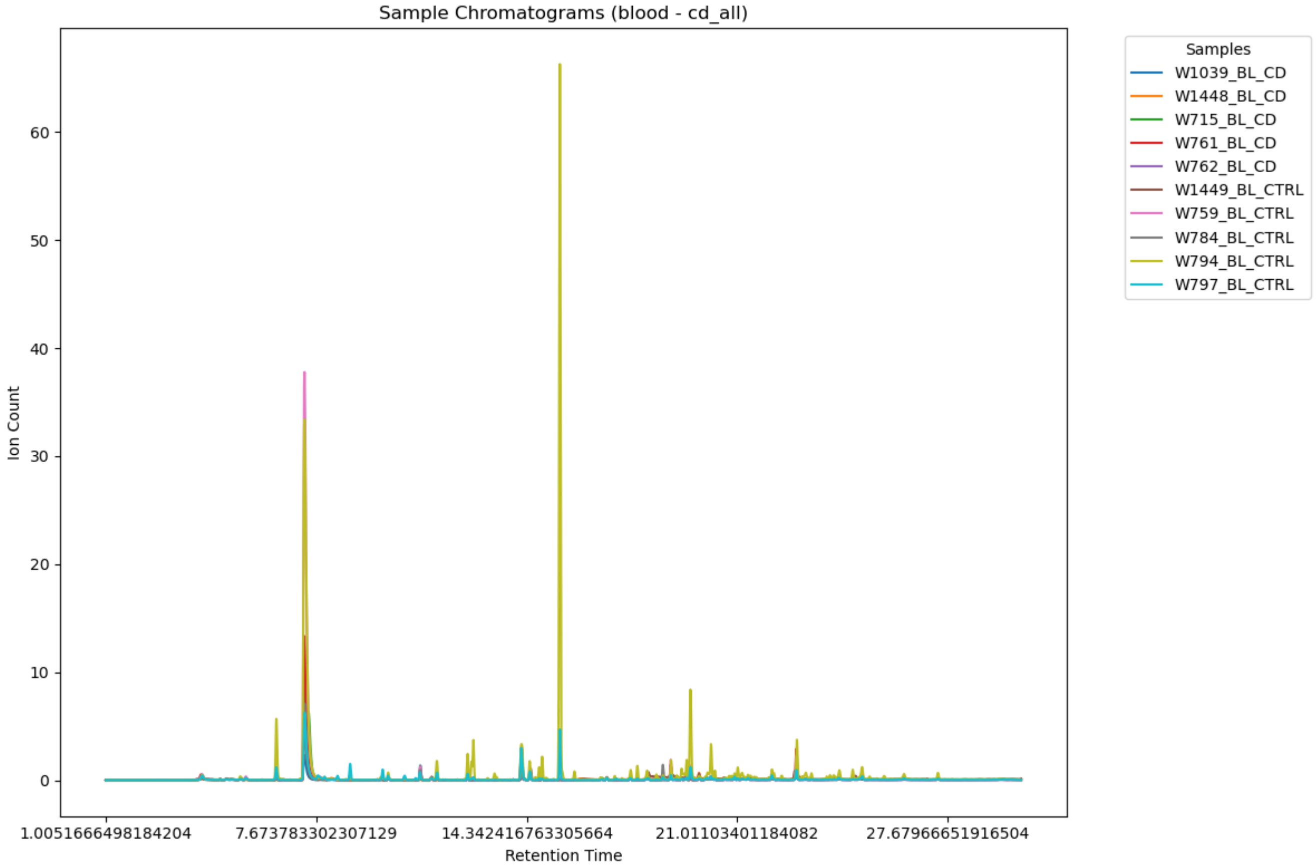


Figure 1: Chromatogram for blood sample in cd_all dataset.

Figure 2 depicts the chromatograms of blood samples from the CD_Ctrl dataset. A similar pattern emerges here, but the intensity of the peaks varies dramatically across CD and control samples, particularly at the strong peaks about 7.67 minutes. The peaks in the control samples are less prominent than those in the CD samples, suggesting that specific metabolites may be higher in Crohn's disease patients.

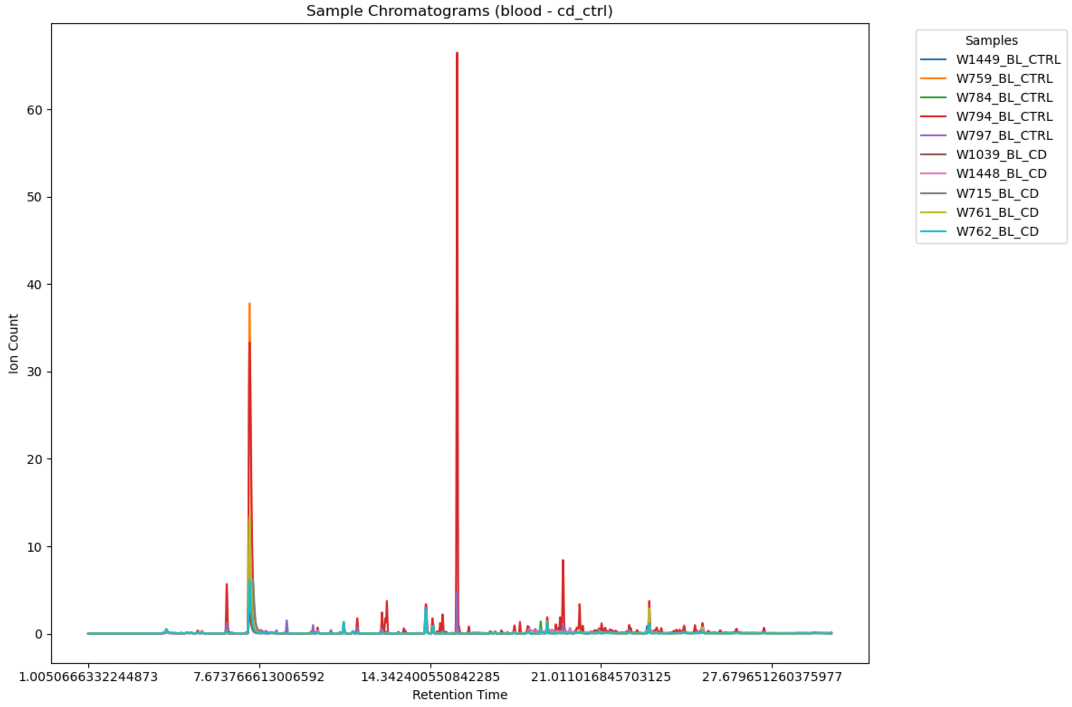


Figure 2: Chromatogram for blood sample in cd_ctrl dataset

These chromatograms depict possible metabolic changes in blood that could be exploited for diagnostic purposes. However, the similarity in peak patterns between CD and control samples suggests that blood may not provide the most obvious metabolic signature for distinguishing CD from control conditions.

Classification Analysis

The classification performance of the SVM and Random Forest models on blood samples illustrates the sample's diagnostic potential.

- **SVM Accuracy**: For the CD_All dataset, the SVM has an accuracy of 0.7544, indicating average classification performance. However, with the CD_Ctrl dataset, the SVM accuracy dropped to 0.5938, demonstrating that distinguishing between CD and control using blood samples is more challenging.

Due to formatting reasons, Figure 3 is displayed on the next page.

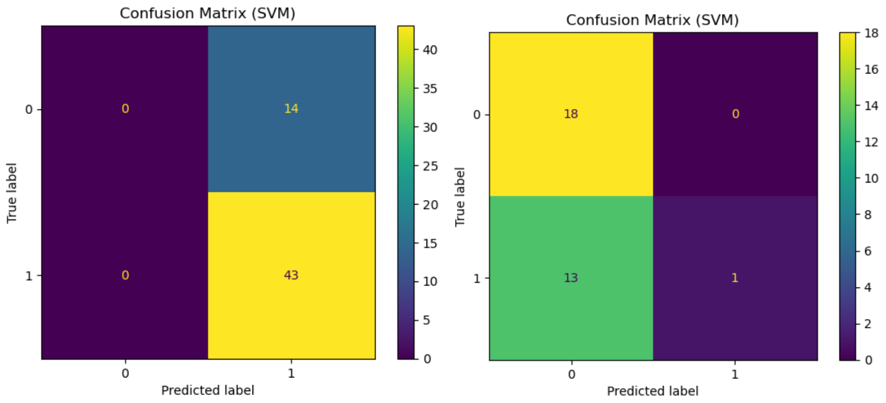


Figure 3: SVM Confusion Matrixes

Description: Left matrix is for the cd_all dataset, with an accuracy of 0.7544. Right matrix is for the cd_ctrl dataset, with an accuracy of 0.5938.

- **Random Forest Accuracy**: The Random Forest classifier had perfect accuracy (1.0) on both the CD_All and CD_Ctrl datasets. This could indicate overfitting, particularly if the model is extremely sophisticated in relation to the dataset size, or it could reveal a major pattern in the data that the SVM was unable to capture correctly.

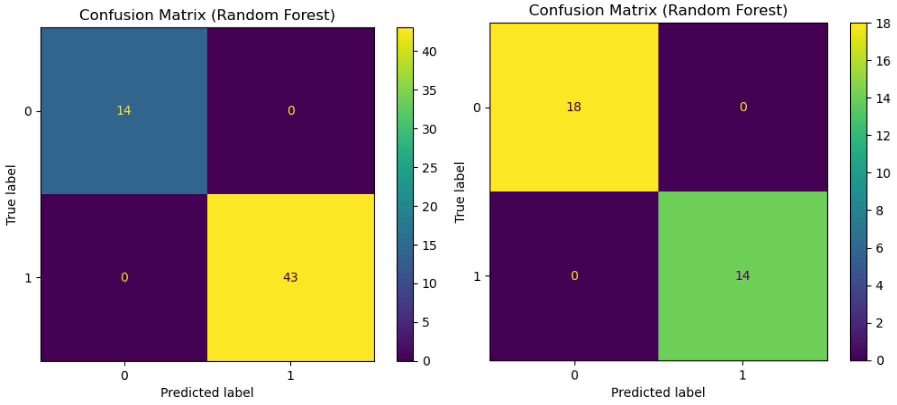


Figure 4: RF Confusion Matrixes

Description: Left matrix is for the cd_all dataset, with an accuracy of 1.0. Right matrix is for the cd_ctrl dataset, with an accuracy of 1.0

Bootstrap and Permutation Test Results

Bootstrap and permutation tests were used to assess the classifiers' robustness.

Dataset	SVM Bootstrap	Random Forest Bootstrap	SVM Permutation	Random Forest Permutation
CD_All	0.735	0.6944	0.7489	0.7111
CD_Ctrl	0.455	0.459	0.486	0.467

Table 1: Bootstrap and Permutation Test Results for SVM and Random Forest Models on Blood Samples (CD_All and CD_Ctrl Datasets).

The bootstrap and permutation results for the CD_All dataset show that, while the SVM performed quite consistently, the Random Forest's high accuracy may become less reliable with resampling. The permutation tests, which randomise labels to establish the significance of the results, demonstrate that the reported accuracies are statistically significant, although they raise concerns about overfitting in the Random Forest.

In comparison, the CD_Ctrl dataset displays lower average accuracies across all bootstrap and permutation tests for both SVM and Random Forest models. This highlights the difficulty of distinguishing between Crohn's disease and normal samples using solely blood biomarkers.

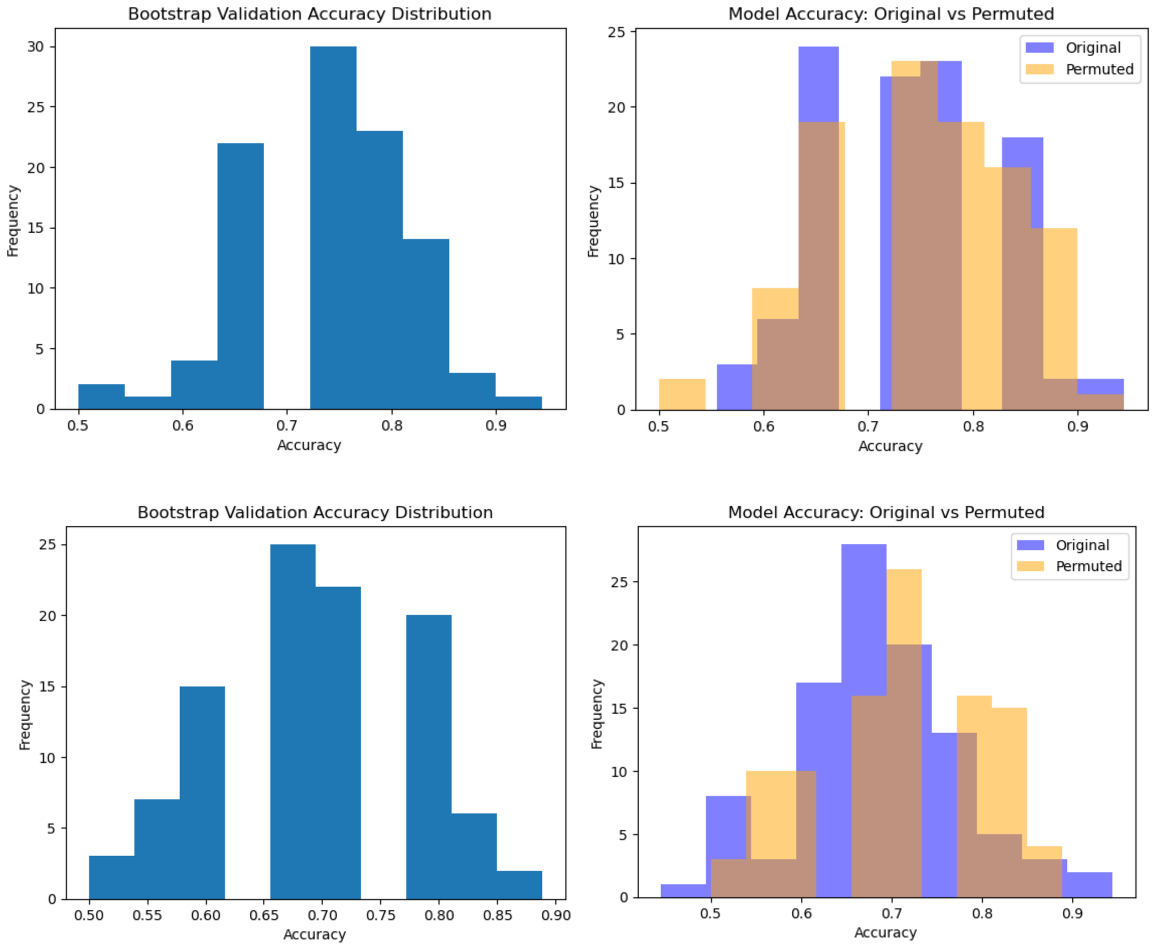


Figure 4: Bootstrap Validation and Permutation Testing for SVM and Random Forest on CD_All Blood Sample Dataset.
Description: The top graphs represent the SVM model results, while the bottom graphs illustrate the outcomes from the Random Forest model.

Due to formatting reasons, Figure 5 is displayed on the next page.

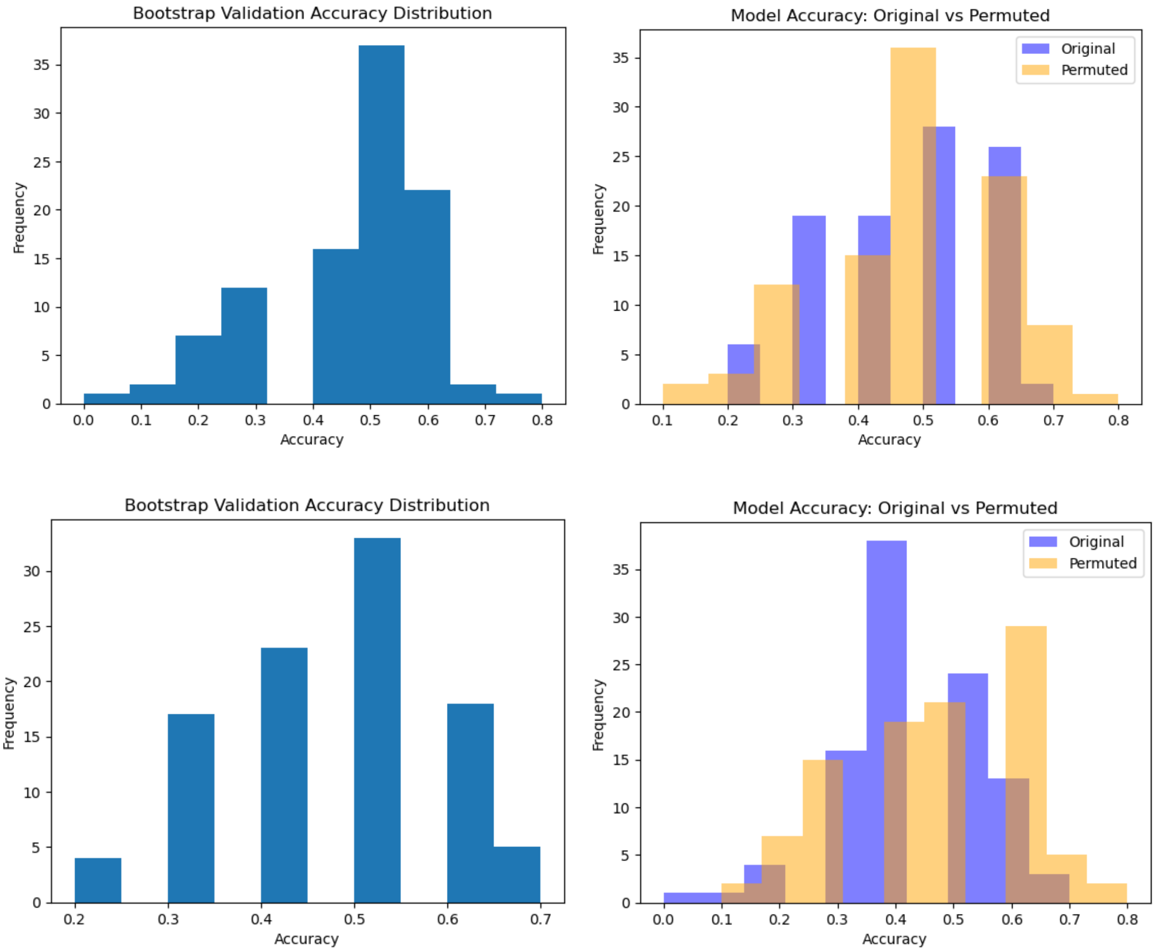


Figure 5: Bootstrap Validation and Permutation Testing for SVM and Random Forest on CD_Ctrl Blood Sample Dataset.
Description: The top graphs represent the SVM model results, while the bottom graphs illustrate the outcomes from the Random Forest model.

DISCUSSION: BLOOD SAMPLES

The research of blood samples found moderate efficiency in distinguishing Crohn's Disease (CD) from control conditions. While the chromatograms revealed changes in metabolite profiles, the overall comparability of CD and control samples suggested that blood may not be the most distinguishing metabolic signature for detecting Crohn's disease. This discovery is consistent with previous research, which has shown that blood biomarkers, while indicative of inflammatory processes, may lack the specificity required for accurate Crohn's disease diagnosis when used alone (Dunn, et al., 2011).

The Support Vector Machine (SVM) model achieved an accuracy of 0.7544 on the CD_All dataset, indicating that blood samples have diagnostic potential. However, the lower accuracy of 0.5938 on the CD_Ctrl dataset demonstrates the problems associated with relying solely on blood for diagnosis. The Random Forest model's 100% accuracy may indicate overfitting, particularly given the lower bootstrap validation accuracies (0.6944 for CD_All and 0.459 for CD_Ctrl). This mismatch demonstrates the model's potential overfitting to the training data, which is a common problem in advanced models applied to limited datasets (Hastie, et al., 2009).

The findings imply that, while blood biomarkers can aid in the detection of Crohn's disease, they may not be unique enough to be used as standalone indicators. Other biological samples, such as urine or faeces, may provide more distinct metabolic fingerprints, improving diagnostic accuracy when combined with blood biomarkers (Aldars-Garcia, et al., 2021).

While the analysis of blood samples gave crucial insights into the metabolomic patterns associated with Crohn's disease, various shortcomings were discovered during the inquiry that must be carefully addressed. These limitations underscore not only the inherent difficulty of using blood as a single diagnostic tool, but also the complexities of accurately identifying disease-specific biomarkers.

1. **Overfitting in the Random Forest Model**: The Random Forest model's observed 100% accuracy indicates potential overfitting, which could imperil the model's capacity to generalise to new data. To address this issue, more robust cross-validation processes or a reduction in model design may be required to reduce overfitting and increase generalisability (Hastie, et al., 2009).
2. **Similarity in Chromatograms**: The striking similarity in chromatograms between Crohn's Disease (CD) and control groups implies that blood samples may not fully reflect the metabolomic changes associated with CD. This constraint highlights the need to study additional biomarkers or mix multiple biological sample types in order to obtain more specific and trustworthy diagnostic results (Aldars-Garcia, et al., 2021).
3. **Challenges in Diagnostic Accuracy**: The reduced accuracy in the CD_Ctrl dataset suggests that blood biomarkers alone may not be sufficient for diagnosing Crohn's disease. This limitation highlights the importance of employing a multi-sample diagnostic technique or incorporating new biomarkers to improve the accuracy and specificity of Crohn's disease diagnosis (Aldars-Garcia, et al., 2021).

BREATH SAMPLES

Chromatogram Analysis

Chromatograms were constructed for both the CD_All and CD_Ctrl datasets to provide a visual representation of the metabolite profiles in the breath of persons with Crohn's Disease (CD) versus control subjects. Figure 6 depicts the chromatograms for the breath samples from the CD_All dataset. The retention times and ion counts reflect the samples' metabolomic landscape. Both the CD and control groups show notable peaks, such as those at 7.67 and 14.34 minutes. The intensity and distribution of these peaks differ somewhat between the two groups, implying potential metabolic differences.

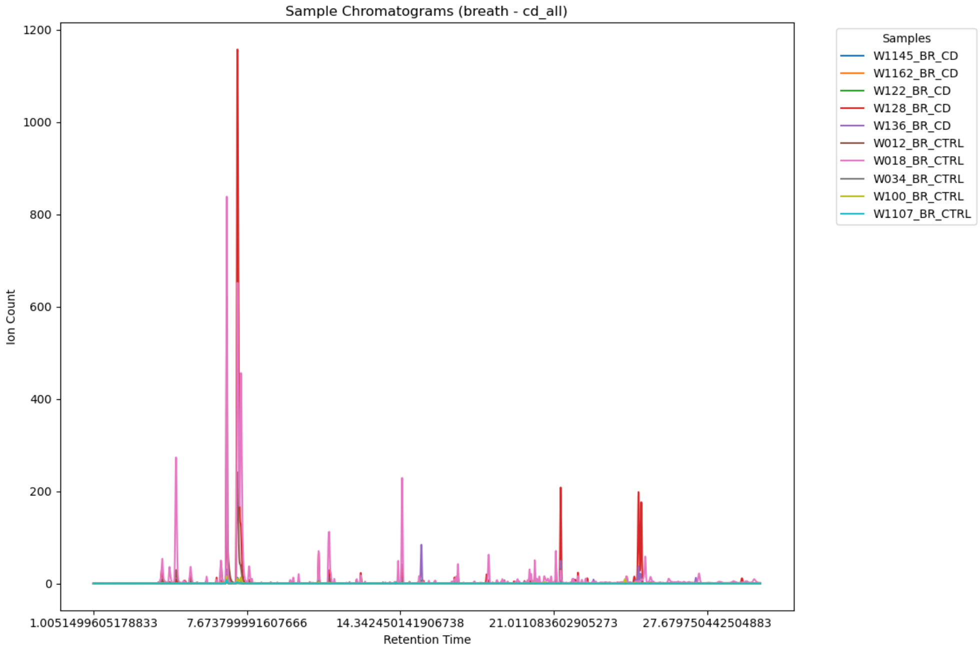


Figure 6: Chromatogram for breath sample in cd_all dataset.

Figure 7 on the other hand, shows the chromatograms for the breath samples from the CD_Ctrl dataset. The chromatograms demonstrate a slightly more noticeable difference in metabolite profiles between the CD and control samples, especially at 7.67 minutes. The control samples exhibit peaks of varied strengths as compared to the CD samples, indicating that breath can identify unique metabolomic changes associated with Crohn's disease.

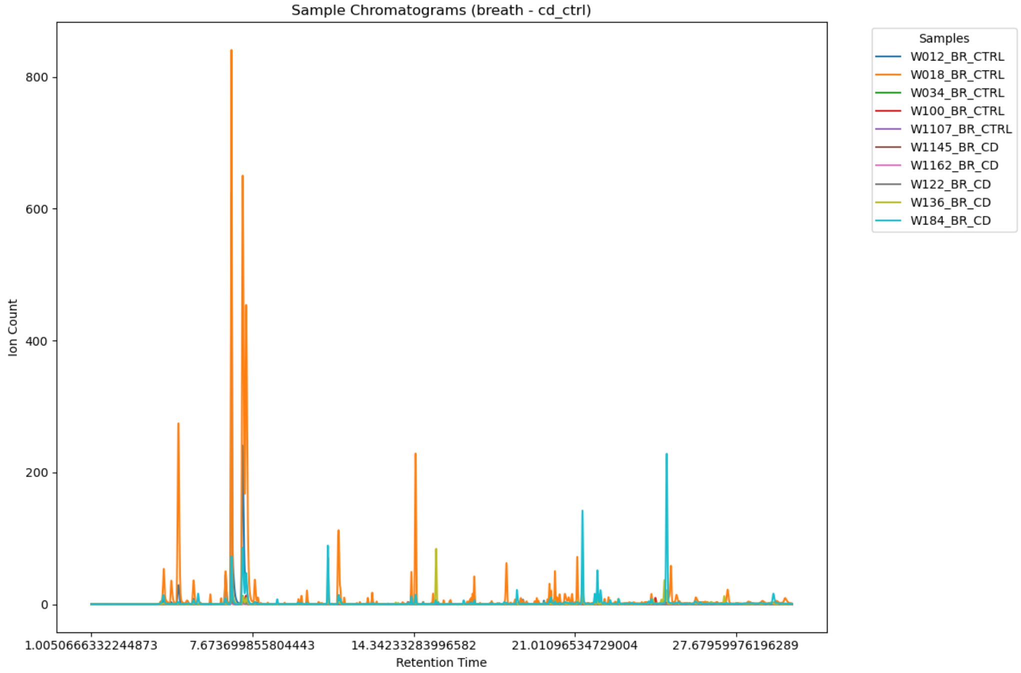


Figure 7: Chromatogram for blood sample in cd_ctrl dataset

When the breath samples are compared side by side, they show the metabolic changes associated with Crohn's disease; however, the differences between CD and control samples are not as obvious as they could be for a solo diagnostic tool.

<u>Classification Analysis</u>

The classification performance of the SVM and Random Forest models on breath samples illustrates the sample's diagnostic potential.

- **SVM Accuracy**: For the CD_All dataset, the SVM has an accuracy of 0.7333, which suggests a moderate performance when it comes to distinguishing between CD and control samples. CD_Ctrl dataset indicates a greater difficulty in differentiating these groups using breath samples alone as its accuracy is at 0.5714

Due to formatting reasons, Figure 8 is displayed on the next page.

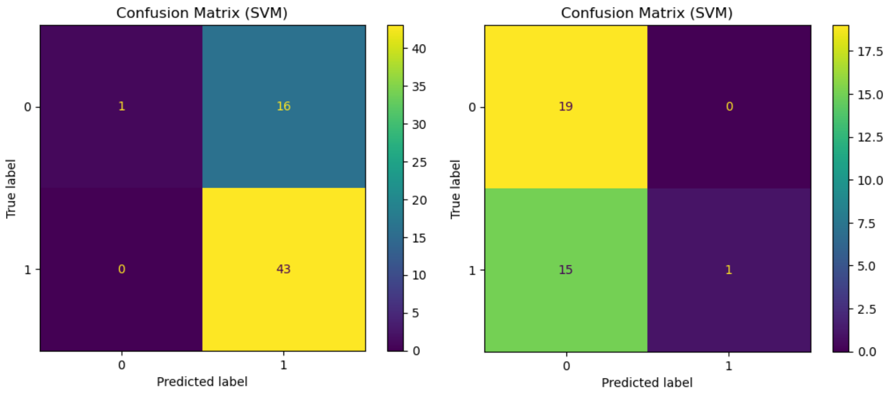


Figure 8: SVM Confusion Matrixes

Description: Left matrix is for the cd_all dataset, with an accuracy of 0.7333. Right matrix is for the cd_ctrl dataset, with an accuracy of 0.5714.

- **Random Forest Accuracy**: Random Forest model, on both datasets – CD_All and CD_Ctrl – are at 1.0, which leans towards overfitting.

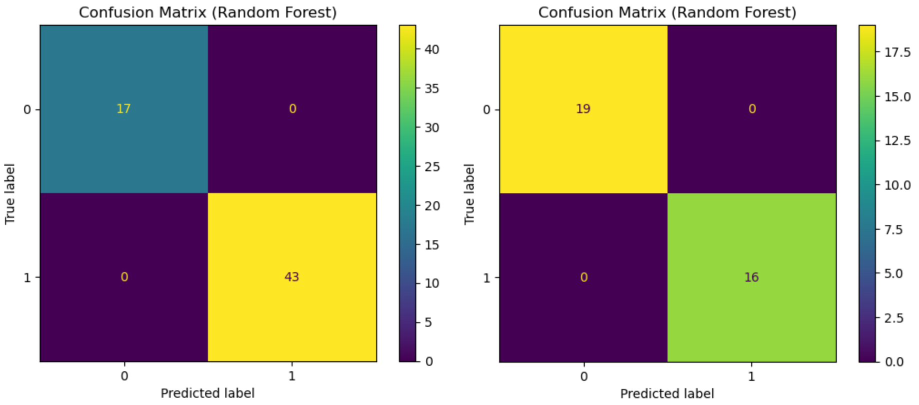


Figure 9: RF Confusion Matrixes

Description: Left matrix is for the cd_all dataset, with an accuracy of 1.0. Right matrix is for the cd_ctrl dataset, with an accuracy of 1.0

Bootstrap and Permutation Test Results

Bootstrap and permutation tests were used to assess the classifiers' robustness.

Dataset	SVM Bootstrap	Random Forest Bootstrap	SVM Permutation	Random Forest Permutation
CD_All	0.6972	0.7439	0.7133	0.6372
CD_Ctrl	0.4218	0.6373	0.44	0.4536

Table 2: Bootstrap and Permutation Test Results for SVM and Random Forest Models on Breath Samples (CD_All and CD_Ctrl Datasets).

The bootstrap and permutation results for the CD_All dataset show that, although the model is rather stable, the Random Forest's flawless accuracy may cause issues when resampled. The permutation tests demonstrate that the observed accuracies are statistically significant, but also raise concerns about overfitting – as mentioned above – in the Random Forest model.

The lower bootstrap and permutation accuracies for both models in the CD_Ctrl dataset demonstrate the difficulty in discriminating between Crohn's disease and control participants using only breath samples.

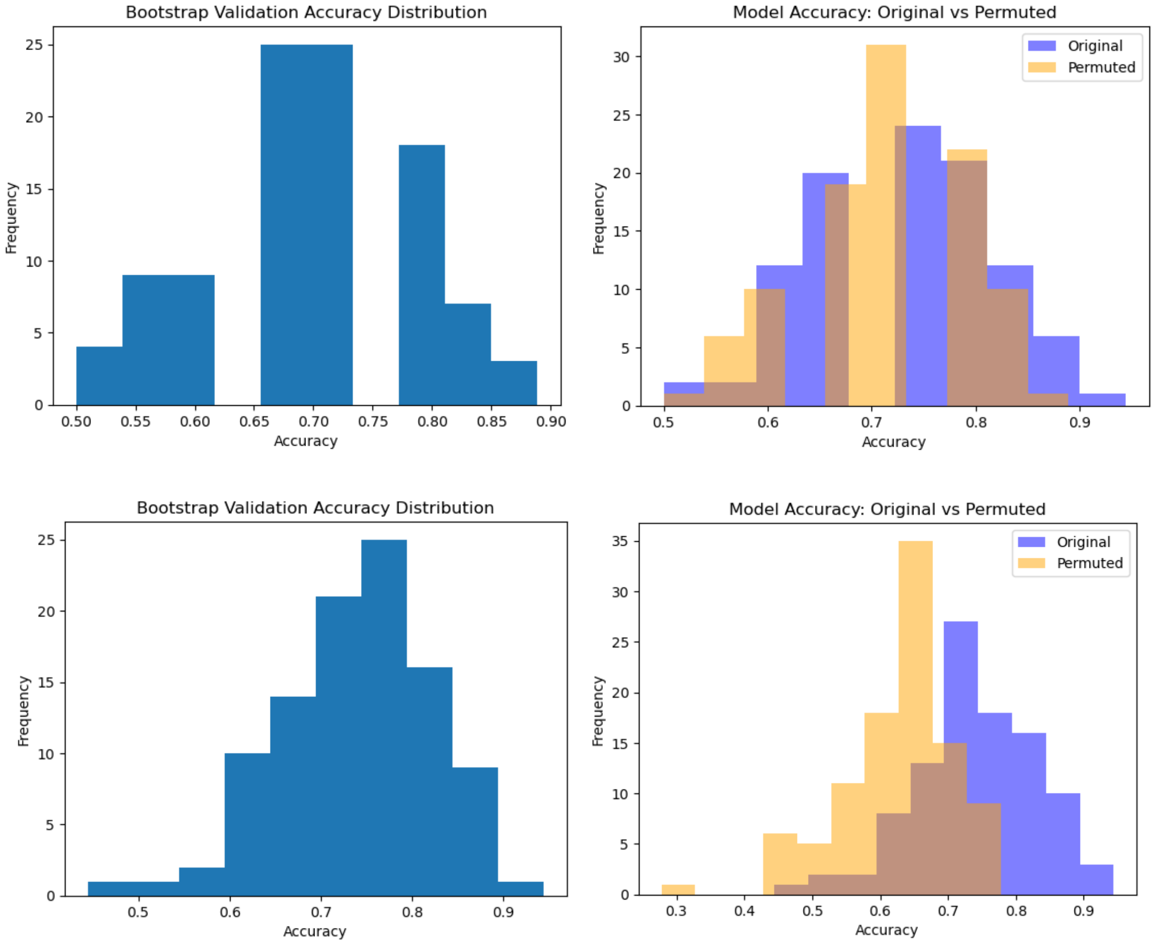


Figure 10: Bootstrap Validation and Permutation Testing for SVM and Random Forest on CD_All Breath Sample Dataset. *Description: The top graphs represent the SVM model results, while the bottom graphs illustrate the outcomes from the Random Forest model.*

Due to formatting reasons, Figure 11 is displayed on the next page.

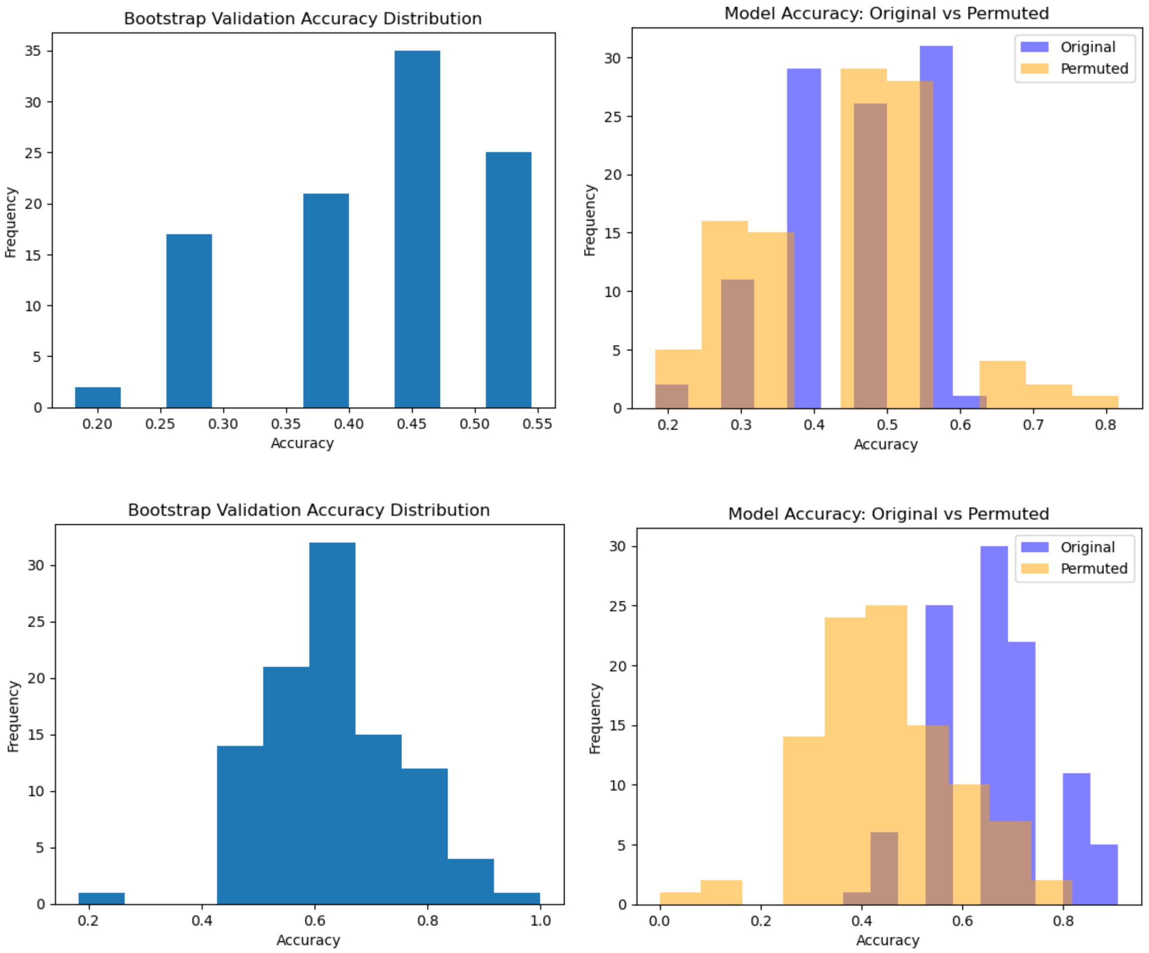


Figure 11: Bootstrap Validation and Permutation Testing for SVM and Random Forest on CD_Ctrl Breath Sample Dataset.
Description: The top graphs represent the SVM model results, while the bottom graphs illustrate the outcomes from the Random Forest model.

DISCUSSION: BREATH SAMPLES

Breath sample analysis has demonstrated substantial efficacy in differentiating Crohn's Disease from control conditions. While the chromatograms showed changes in metabolite profiles between CD and control samples, the overlap shows that breath may not be the most reliable metabolic signature for Crohn's disease diagnosis.

The SVM model's accuracy of 0.7333 on the CD_All dataset shows that breath samples may be useful for diagnostic purposes. However, the CD_Ctrl dataset's lower accuracy (0.5714) demonstrates the difficulties associated with relying solely on breath samples for reliable differentiation. Furthermore, the Random Forest model's immaculate accuracy across both datasets raises concerns about possible overfitting, as seen by significantly lower bootstrap validation accuracies (0.7439 for CD_All and 0.6373 for CD_Ctrl). This problem has been documented in similar scenarios, where models exhibit high accuracy due to overfitting yet fail to generalise efficiently to new data (Hastie, et al., 2009).

While breath analysis can detect particular metabolic changes associated with Crohn's disease, it may not provide enough specificity to be used as a single diagnostic tool (Aldars-Garcia, et al., 2021). The Breath Samples are subject to overfitting in the Random Forest model, just like the Blood Samples. The chromatograms of Crohn's disease and control samples overlap, indicating that breath analysis cannot capture the many metabolomic changes, and the lower accuracies in the CD_Ctrl dataset suggest that breath samples alone may not be reliable enough to diagnose Crohn's disease.

FAECAL SAMPLES

Chromatogram Analysis

The faecal samples were treated similarly, and chromatograms were generated for both the CD_All and CD_Ctrl datasets. Figure 12 depicts the retention periods and ion counts, indicating a diverse spectrum of metabolites in the samples. Peaks appear at retention periods of 7.67 minutes, 14.34 minutes, and 21.01 minutes, with significant differences in their intensity and presence between the CD and controls.

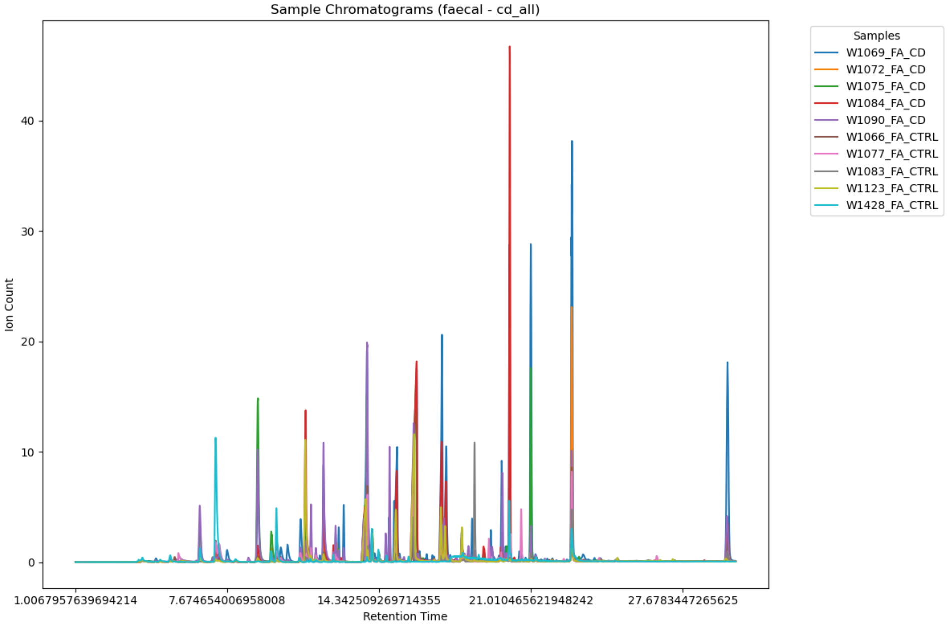


Figure 12: Chromatogram for faecal sample in cd_all dataset.

Figure 13 displays a pattern similar to that seen in the CD_All dataset, but with significantly different intensities. Prominent peaks, particularly at 7.67 and 21.01 minutes, present in both CD and control samples, but their intensities differ, showing that metabolites are higher in Crohn's disease than controls.

Due to formatting reasons, Figure 13 is displayed on the next page.

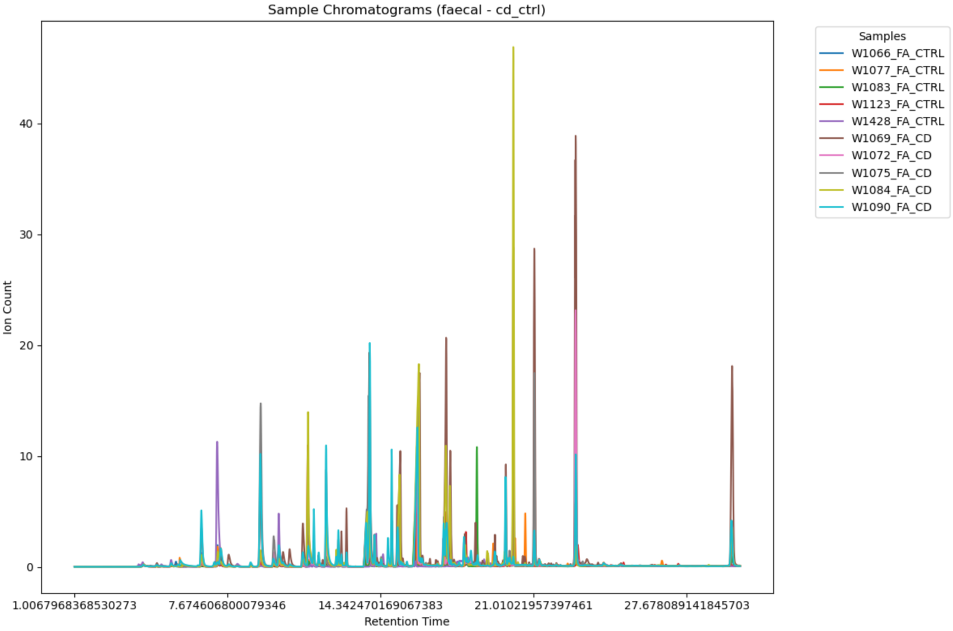


Figure 13: Chromatogram for faecal sample in cd_ctrl dataset

Faecal samples, rather than blood and breath samples, tend to provide a more reliable source of metabolic information for distinguishing Crohn's disease from control conditions. The observed differences in peak patterns between the CD and control groups indicate that faecal samples can detect various metabolic fingerprints associated with Crohn's disease.

Classification Analysis

The classification performance of the SVM and Random Forest models on faecal samples illustrates the sample's diagnostic potential.

- **SVM Accuracy**: The CD_All dataset achieved an accuracy of 0.7778, showing that the SVM model performed reasonably well in distinguishing Crohn's Disease and control samples throughout the whole dataset. The model's ability to achieve more than 77% accuracy implies that stool samples contain useful diagnostic information, while there is still room for improvement, particularly in detecting sickness nuances. On the other hand, the CD_Ctrl dataset trained on the SVM model achieves a higher accuracy of approximately 95.65%, indicating that when the model is focused on distinguishing between CD and a more narrowly defined control group, it performs exceptionally well, capturing the key metabolic differences that distinguish Crohn's disease from healthy states.

Due to formatting reasons, Figure 14 is displayed on the next page.

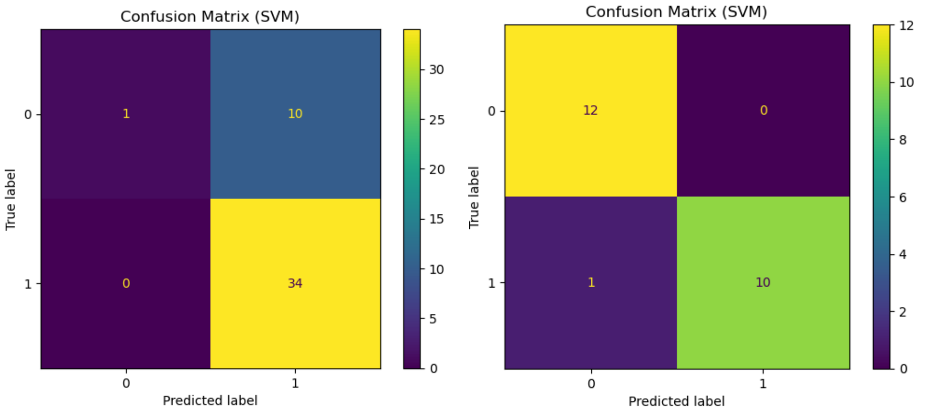


Figure 14: SVM Confusion Matrixes

Description: Left matrix is for the cd_all dataset, with an accuracy of 0.7778. Right matrix is for the cd_ctrl dataset, with an accuracy of 0.9565.

- **Random Forest Accuracy**: Once again, the Random Forest model obtains an exact accuracy of 1.0 on both the CD_All and CD_Ctrl datasets, indicating overfitting. While this implies that the models are capable of capturing strong patterns in the data, there is a risk that they will not generalise effectively to new, unknown data, resulting in overfitting as the model becomes too complicated, capturing not only the underlying pattern but also the noise in training data.

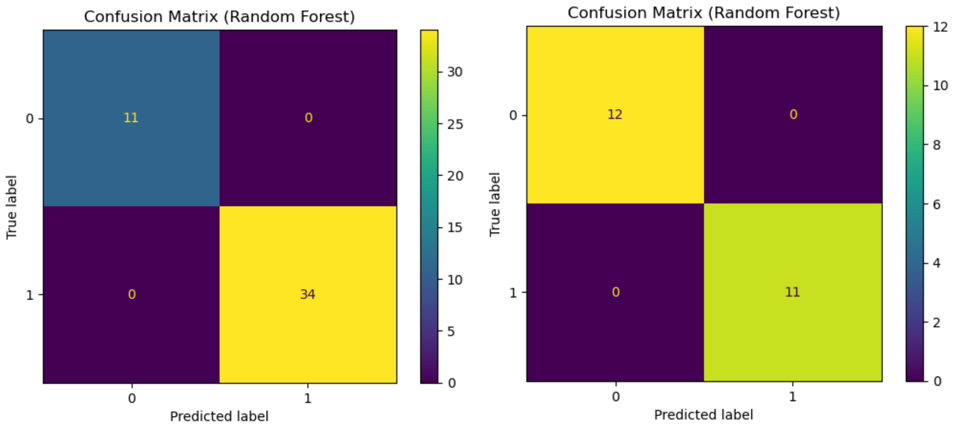


Figure 15: RF Confusion Matrixes

Description: Left matrix is for the cd_all dataset, with an accuracy of 1.0. Right matrix is for the cd_ctrl dataset, with an accuracy of 1.0

Bootstrap and Permutation Test Results

Bootstrap and permutation tests were used to assess the classifiers' robustness.

Dataset	SVM Bootstrap	Random Forest Bootstrap	SVM Permutation	Random Forest Permutation
CD_All	0.7536	0.8007	0.7550	0.6707
CD_Ctrl	0.6714	0.8571	0.3900	0.4871

Table 3: Bootstrap and Permutation Test Results for SVM and Random Forest Models on Faecal Samples (CD_All and CD_Ctrl Datasets).

The SVM model performs rather consistently on the CD_All dataset, with a bootstrap average accuracy of 0.7536, which is slightly lower than the original accuracy. This indicates that the SVM model is typically robust, but resampling causes a slight performance reduction. The SVM model's permutation test accuracy of 0.7550 indicates that the model's ability to distinguish between Crohn's Disease (CD) and control samples is not attributable to random fluctuations in the data, hence improving the SVM's performance reliability.

Despite having an initial accuracy of 100%, the Random Forest model's bootstrap and permutation test accuracies drop to 0.8007 and 0.6707, respectively. The decrease in accuracy observed during bootstrapping suggests that the model may have overfitted the data, capturing patterns that are not fully generalisable across different subsets of the data. The large decline in permutation test accuracy to 67% exacerbates the issue, implying that the previous high accuracy was slightly misleading, with genuine generalisable accuracy closer to 67-80%.

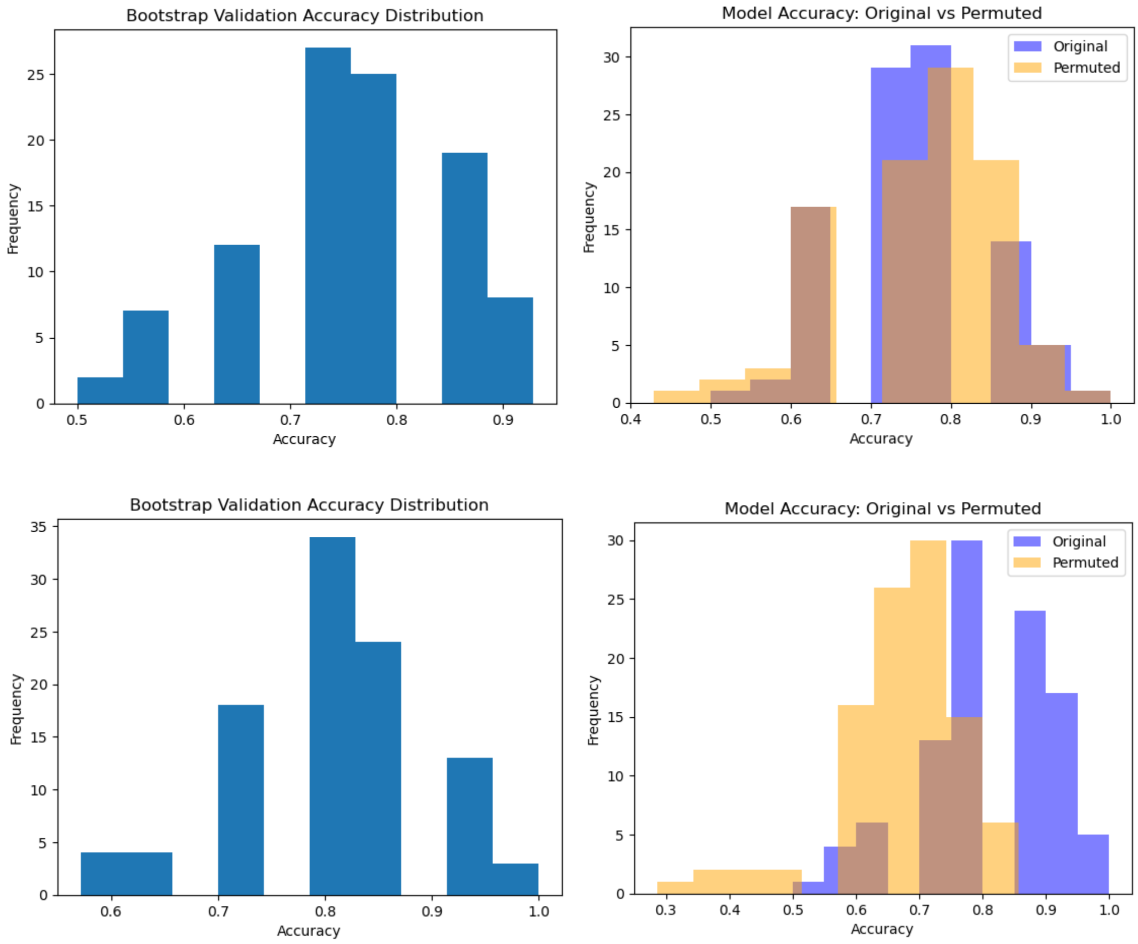


Figure 16: Bootstrap Validation and Permutation Testing for SVM and Random Forest on CD_All Faecal Sample Dataset.
Description: The top graphs represent the SVM model results, while the bottom graphs illustrate the outcomes from the Random Forest model.

The CD_Ctrl dataset shows a more significant performance reduction, particularly for the SVM model. The bootstrap average accuracy drops to 0.6714, which is a significant decrease from the original accuracy and likely instability in the model's performance across multiple data resampling. This variability suggests that the SVM model may have trouble distinguishing between CD and control samples in the dataset. The Random Forest model, albeit having a higher bootstrap average accuracy of 0.8571, exhibits a decrease when compared to its initial perfect accuracy. This illustrates that, while the model detects strong patterns in the data, these patterns may not be entirely generalisable, solidifying the overfitting problem. The permutation test results support this, with the SVM model's accuracy dropping dramatically to 0.3900 and the Random Forest model's accuracy lowering to 0.4871. These reductions suggest that the previously recorded high accuracies were exaggerated, and that the models' true discriminatory power, particularly the SVM model, may be lower than previously anticipated.

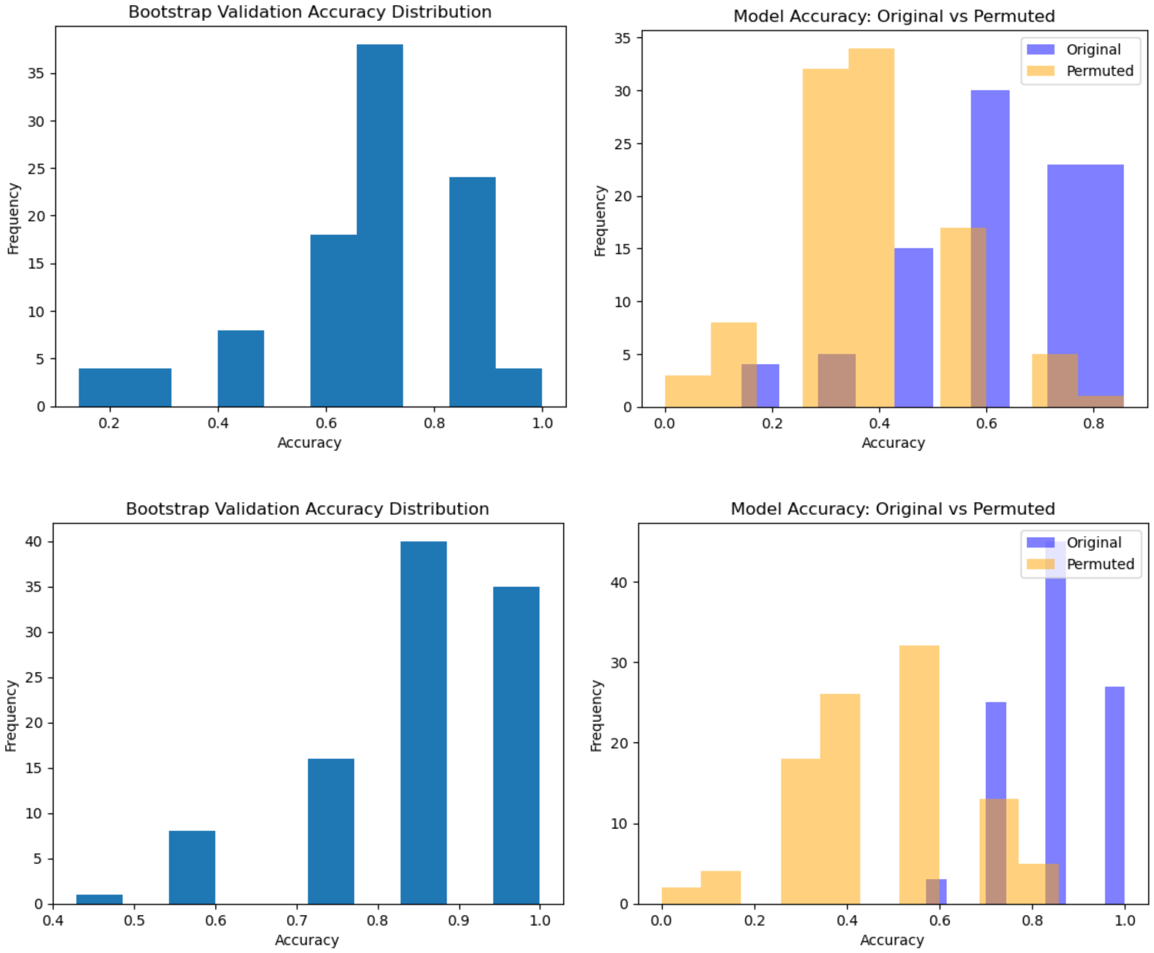


Figure 17: Bootstrap Validation and Permutation Testing for SVM and Random Forest on CD_Ctrl Faecal Sample Dataset.
Description: The top graphs represent the SVM model results, while the bottom graphs illustrate the outcomes from the Random Forest model.

DISCUSSION: FAECAL SAMPLES

The stool samples examined in this section indicated significant metabolic variations between Crohn's disease (CD) patients and controls. The chromatograms for the CD_All and CD_Ctrl datasets showed significant peaks at retention durations of 7.67, 14.34, and 21.01 minutes, respectively. These peaks, which were consistently greater in CD samples, suggest an increase in metabolites linked with Crohn's disease inflammation. Faecal metabolism has been shown to provide valuable insights into gastrointestinal disorders, particularly in the discovery of possible biomarkers for conditions such as Crohn's disease **(Vila, et al., 2023)**.

The different biochemical markers found in faeces highlight this sample type's potential for identifying Crohn's disease. Compared to blood or breath tests, faecal samples appear to be a more trustworthy source of diagnostic information, because the metabolites detected in faeces more directly reflect the gut environment where Crohn's Disease exerts its principal effects (Vila, et al., 2023). This is especially relevant because the gut microbiota and its associated metabolic products are increasingly identified as critical components in the pathogenesis of Crohn's disease (Ma, et al., 2022).

The classification models applied on these stool samples offered more evidence of their diagnostic utility. The SVM model achieved a reasonable accuracy of 0.7778 on the CD_All dataset, demonstrating that it was capable of distinguishing between CD and control samples over a wide dataset. However, when trained on the CD_Ctrl dataset, the SVM model performed substantially better, with an accuracy of 0.9565. This improved accuracy shows that the SVM model, when applied to a more strictly described dataset, can better capture major metabolic differences between Crohn's disease and healthy controls. The Random Forest model initially achieved 100% accuracy on both the CD_All and CD_Ctrl datasets, indicating that it is a powerful diagnostic tool. However, the model's performance dropped drastically throughout the bootstrap and permutation tests, indicating the risk of overfitting. The observed decrease in accuracy to around 67% for the CD_All dataset during permutation testing suggests that the model may be unreliable when applied to new, previously unseen data (Hastie, et al., 2009). This emphasises the significance of exercising caution when interpreting these findings, since the Random Forest model can detect trends in training data but may also accumulate noise, limiting its generalisability.

While the faecal samples demonstrated potential, there are some limitations to consider; The large decline in accuracy of the Random Forest model during permutation testing reveals that overfitting occurs, as mentioned previous including being observed in earlier samples as well – blood and breath. This overfitting could be reduced by using more stringent cross-validation procedures or changing the model to improve its generalisability. Furthermore, the overlap in chromatograms between CD and control samples, albeit less noticeable than in other sample types, suggests that faecal samples may not fully represent the complexities of Crohn’s disease pathology. As a result, combining faecal metabolomics with other diagnostic techniques, such as multi-omics methodologies or biomarkers from other biological matrices could improve its accuracy (Vila, et al., 2023) (Lloyd-Price, et al., 2019).

URINE SAMPLES

Chromatogram Analysis

Urine samples were handled similarly to other biological matrices, and chromatograms were generated for both the CD_All and CD_Ctrl datasets. Figures 18 and 19 show the retention times and ion counts for each metabolite discovered in the samples. Peaks exist at retention durations of approximately 7.67, 14.34, and 21.01 minutes. The intensity and presence of these peaks differ between CD and control samples, indicating that Crohn's disease-related metabolic inefficiencies can be detected in urine.

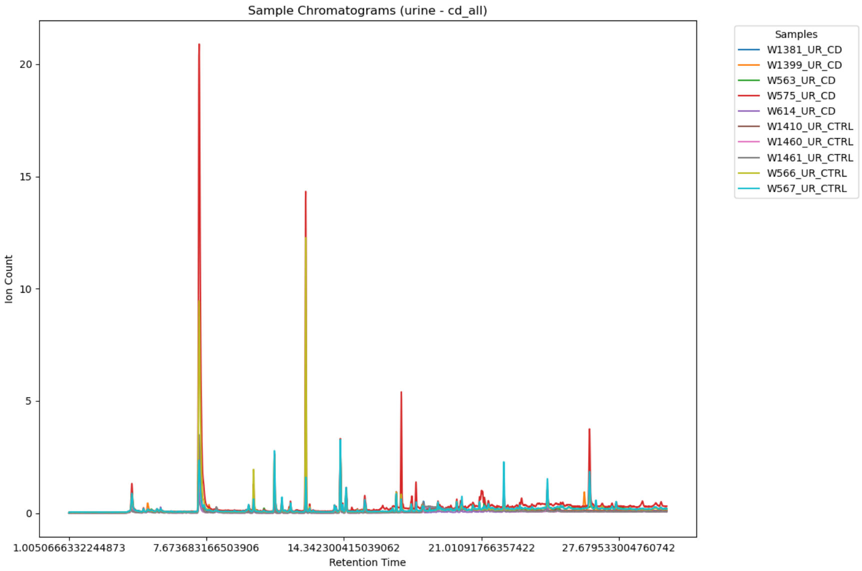


Figure 18: Chromatogram for urine sample in cd_all dataset.

Figure 19 depicts a pattern similar to that found in the CD_All dataset, but with large intensity variability. The CD and control samples both exhibit significant peaks at 7.67 and 14.34 minutes. However, higher intensities in CD samples indicate elevated levels of certain metabolites, which may be associated to Crohn's disease aetiology.

Due to formatting reasons, Figure 19 is displayed on the next page.

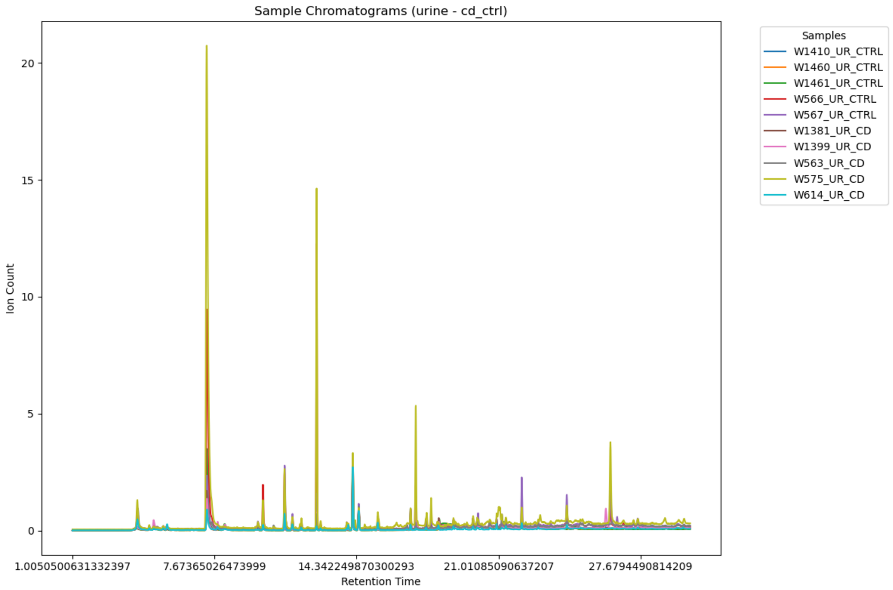


Figure 19: Chromatogram for urine sample in cd_ctrl dataset

Urine samples, like faecal samples, appear to be a useful source of metabolic information for separating Crohn's disease from normal conditions. The distinct peak patterns seen in both the CD and control groups suggest that urine samples can detect metabolic fingerprints associated with Crohn's disease.

Classification Analysis

The classification performance of the SVM and Random Forest models on urine samples illustrates the sample's diagnostic potential.

- **SVM Accuracy**: The CD_All dataset has an accuracy of 0.8056, indicating that the SVM model performed well in distinguishing Crohn's Disease from control samples across the dataset. The model's higher than 80% accuracy indicates that urine samples include useful diagnostic information. However, there is still room for progress, particularly in diagnosing the complexities of Crohn's disease. On the CD_Ctrl dataset, however, the SVM model had an accuracy of 0.7272. This lower accuracy underlines the model's limitations when applied to a more narrowly defined control group, indicating that urine samples alone may not be adequate for meaningful diagnosis in the absence of other markers.

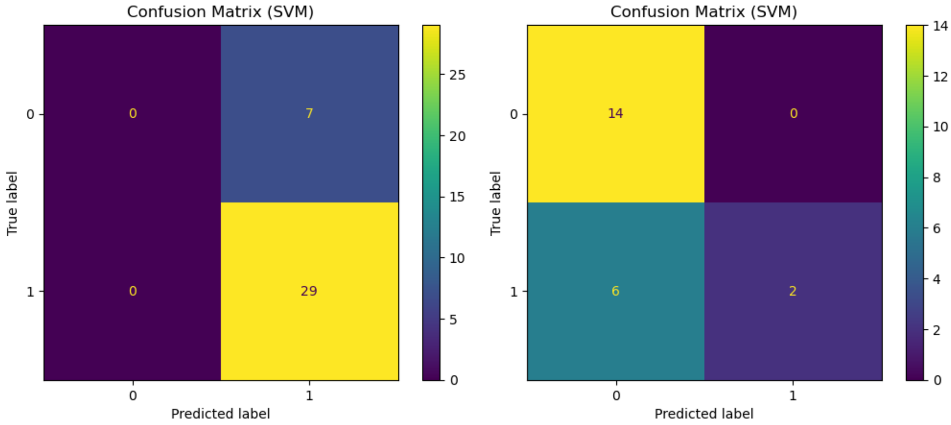
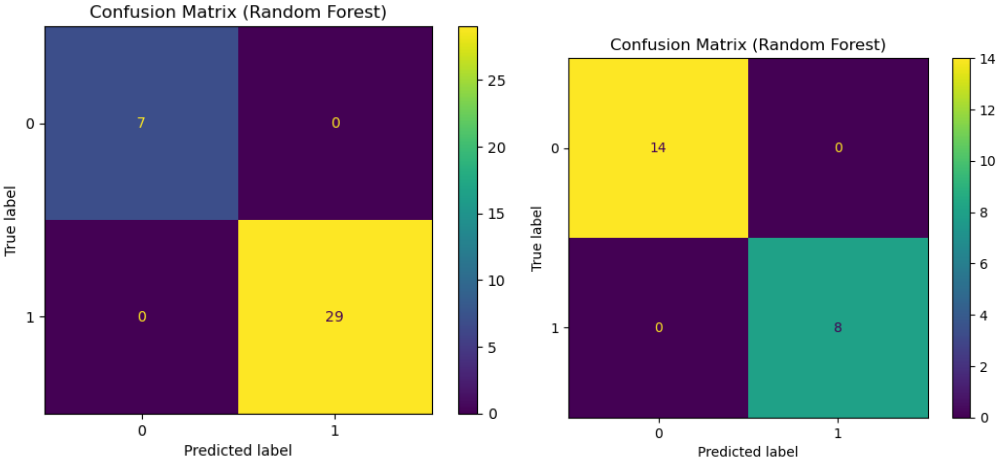


Figure 20: SVM Confusion Matrixes

Description: Left matrix is for the cd_all dataset, with an accuracy of 0.8055. Right matrix is for the cd_ctrl dataset, with an accuracy of 0.7272.

- **Random Forest Accuracy**: The Random Forest model, like the other sample types, obtained 1.0 accuracy on the CD_All and CD_Ctrl datasets. This means that the model identified significant patterns in the training data. However, the risk of overfitting is obvious: the model may have learnt the training data too well, resulting in noise and irrelevant patterns.


Confusion Matrix (Random Forest)

Confusion Matrix (Random Forest)

Figure 21: RF Confusion Matrixes

Description: Left matrix is for the cd_all dataset, with an accuracy of 1.0. Right matrix is for the cd_ctrl dataset, with an accuracy of 1.0

Bootstrap and Permutation Test Results

Bootstrap and permutation tests were used to assess the classifiers' robustness.

Dataset	SVM Bootstrap	Random Forest Bootstrap	SVM Permutation	Random Forest Permutation
CD_All	0.7955	0.7755	0.8136	0.7400
CD_Ctrl	0.6443	0.6229	0.5757	0.5086

Table 4: Bootstrap and Permutation Test Results for SVM and Random Forest Models on Urine Samples (CD_All and CD_Ctrl Datasets).

The SVM model performed consistently on the CD_All dataset, with a bootstrap average accuracy of 0.7955, somewhat lower than the original accuracy. This implies that the SVM model is generally robust; nevertheless, resampling reduces performance. The SVM model's ability to distinguish between CD and control samples is not random, as evidenced by its permutation test accuracy of 0.8136.

The Random Forest model's bootstrap and permutation test accuracies are 0.7755 and 0.74, respectively. These findings suggest that, whereas the Random Forest model initially displayed 100% accuracy, its true generalisability could be closer to 74-78%, emphasising the importance of cautious interpretation.

Due to formatting reasons, Figure 22 is displayed on the next page.

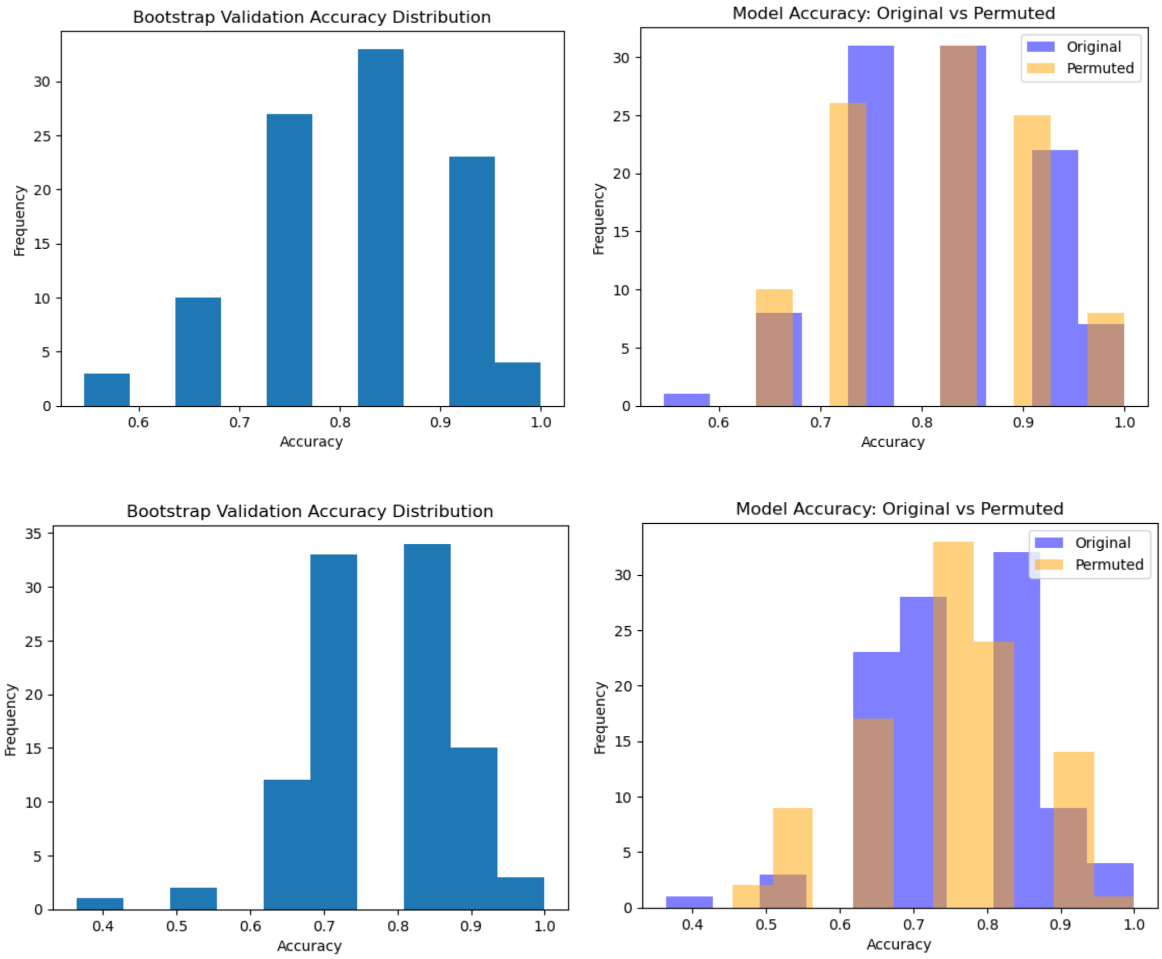


Figure 22: Bootstrap Validation and Permutation Testing for SVM and Random Forest on CD_All Urine Sample Dataset.
Description: The top graphs represent the SVM model results, while the bottom graphs illustrate the outcomes from the Random Forest model.

The CD_Ctrl dataset shows a more dramatic decline in performance, especially for the SVM model. The bootstrap average accuracy falls to 0.6443, revealing a significant variation in model performance between data resamplings. This shows that the SVM model will have difficulty differentiating between CD and control samples in this dataset. Despite a higher bootstrap accuracy of 0.6229, the Random Forest model remains less than 100 percent correct, raising concerns about overfitting. The considerable fall in permutation test accuracy for both models highlights a potential overfitting issue, implying that their genuine discriminatory capacity may be lower than previously anticipated.

Due to formatting reasons, Figure 23 is displayed on the next page.

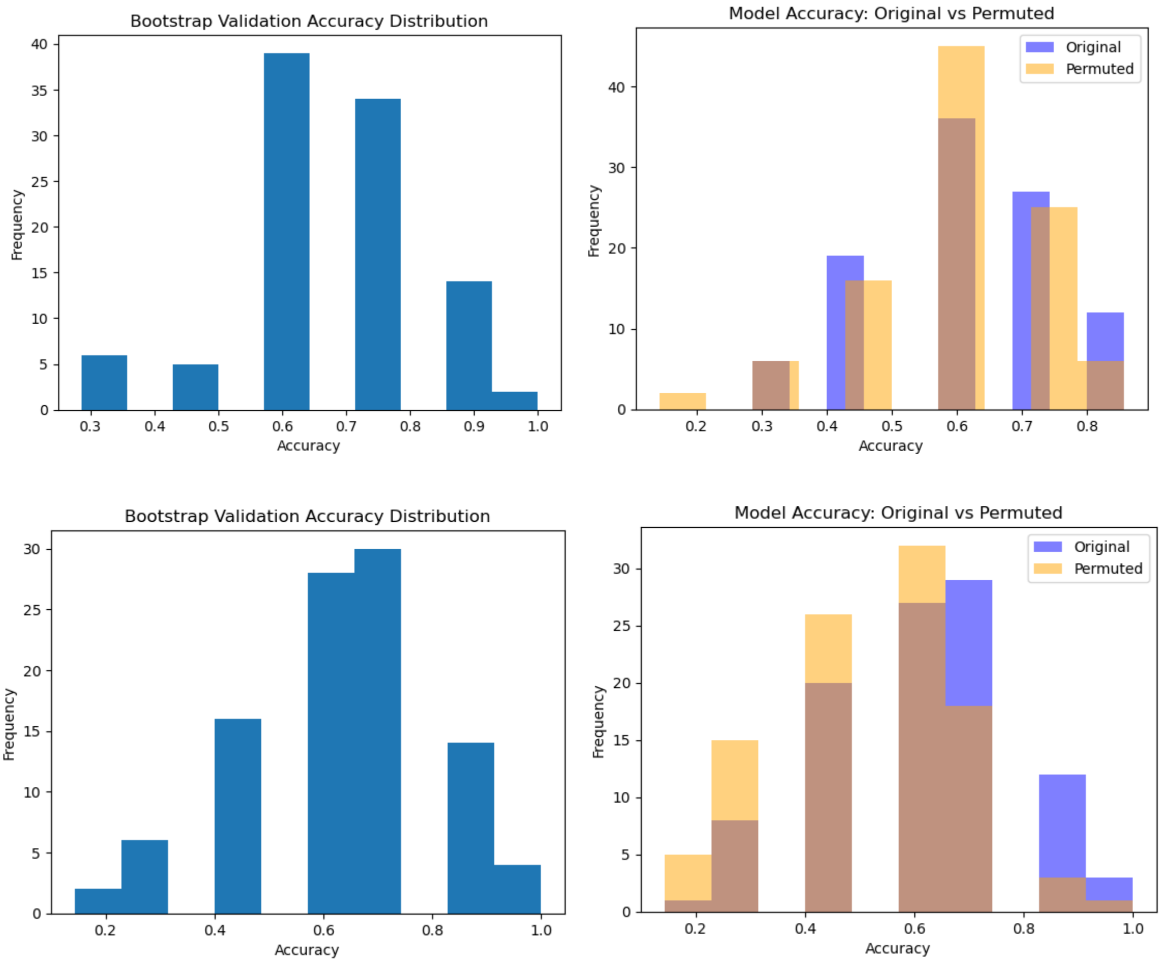


Figure 17: Bootstrap Validation and Permutation Testing for SVM and Random Forest on CD_Ctrl Urine Sample Dataset.
Description: The top graphs represent the SVM model results, while the bottom graphs illustrate the outcomes from the Random Forest model.

DISCUSSION: URINE SAMPLES

Urine sample analysis indicated significant metabolic differences between Crohn's Disease (CD) patients and controls, as seen by chromatograms generated for the CD_All and CD_Ctrl datasets. The chromatograms show strong peaks at retention durations of 7.67, 14.34, and 21.01 minutes, indicating that certain metabolites are continuously prevalent in CD patients' urine. This conclusion is consistent with prior research, which has identified abnormalities in urine metabolism as a characteristic of inflammatory bowel illnesses such as Crohn's Disease. Urine, as a biological matrix, is particularly valuable for capturing the disease's systemic effects because it contains metabolic byproducts from numerous pathways affected by the chronic inflammation associated with CD (Li, et al., 2008).

The differences in peak intensity between CD and control samples, particularly the higher levels of specific metabolites in CD samples, indicate that urine may be a helpful source of biomarkers for detecting Crohn's disease. This lends support to the concept that urine metabolomics can provide a non-invasive and accessible method of monitoring disease activity and treatment response, as well as identifying possible biomarkers for early diagnosis (Sarosiek, et al., 2015).

The classification models utilised with urine samples show diagnostic promise. On the CD_All dataset, the SVM model had an accuracy of 0.8056, indicating that it was effective at distinguishing CD from control samples. This illustrates that urine samples contain valuable diagnostic information, which the SVM model can accurately capture. However, when applied to the CD_Ctrl dataset, the model's accuracy dropped to 0.7273, highlighting the difficulties of discriminating between CD and controls in a more narrowly defined dataset.

The Random Forest model, which had previously performed flawlessly on both datasets, showed signs of overfitting, as seen by a decrease in accuracy during the bootstrap and permutation tests. The bootstrap accuracy for the CD_All dataset dropped to 0.7755, while the permutation test accuracy fell to 0.74. Similarly, the bootstrap accuracy for the CD_Ctrl dataset was 0.6229, but the permutation accuracy was 0.5086, showing that the model may not generalise effectively to new data. These findings are consistent with previous research that has shown the value of urine metabolomics in the diagnosis and monitoring of Crohn's disease. Urine is becoming increasingly used as a biological matrix for studying systemic disorders due to its non-invasive collection and diverse chemical composition. Certain urine metabolites, such as those involved in amino acid and energy metabolism, have been found to change in CD patients, offering information about disease aetiology and prospective treatment targets (Maszka, et al., 2023).

Several restrictions must be addressed. Random Forest models, as observed in the previous samples, are prone to overfitting, which can be mitigated with improved cross-validation or even simpler models. Furthermore, while urine metabolomics has promise, it may not capture all of the key metabolic changes associated with Crohn's disease, particularly those confined to the gut. As a result, integrating urine metabolism with other diagnostic approaches may lead to more information and better diagnostic results.

CONCLUSION

This study sought to establish whether biological sample type—breath, blood, urine, or faeces—provides the best diagnostic signature for Crohn's disease (CD). Using SVM and Random Forest models on all data sources, faeces emerged as the most reliable source for identifying CD patients from healthy controls. In terms of diagnostic accuracy and consistency, the SVM model outperformed all other sample types, especially on the CD_Ctrl dataset. Faecal samples showed a particular metabolic profile associated with CD, and the Random Forest model performed well, albeit with evidence of overfitting in some cases. Urine samples performed well, particularly in the CD_All dataset, but were less effective in the more narrowly defined CD_Ctrl group. While blood and breath samples were valuable, they had less distinct diagnostic fingerprints, poorer accuracies, and were more prone to overfitting, as evidenced by the bootstrap and permutation test results. When compared to random chance, faecal samples demonstrated a clear diagnostic advantage, beating the findings predicted by a model with no diagnostic capacity. These findings indicate that faecal metabolomics could be a powerful non-invasive diagnostic tool for Crohn's disease, with the possibility for further improvement through multi-omics techniques.

REFERENCES

Aldars-Garcia, L., Gisbert, J. P. & Chaparro, M., 2021. Metabolomics Insights into Inflammatory Bowel Disease: A Comprehensive Review. *Pharmaceuticals (Basel),* 14(11), p. 1190.

Baumgart, D. C. & Sandborn, W. J., 2012. Crohn's disease. *Lancet*, 380(9853), pp. 1590-1605.

Dunn, W. B. et al., 2011. Mass appeal: metabolite identification in mass. *Metabolomics,* 7(1), pp. 4-29.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. Springer Science & Business Media: New York.

Li, M. et al., 2008. Symbiotic gut microbes modulate human metabolic phenotypes. *Proceedings of the National Academy of Sciences of the United States of America,* 105(6), pp. 2117-2122.

Lloyd-Price, J. et al., 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature,* 569(7758), pp. 655-662.

Maszka, P. et al., 2023. Metabolomic Footprint of Disrupted Energetics and Amino Acid Metabolism in Neurodegenerative Diseases: Perspectives for Early Diagnosis and Monitoring of Therapy. *Metabolites,* 13(3), p. 369.

Ma, X. et al., 2022. Gut microbiota in the early stage of Crohn's disease has unique characteristics. *Gut Pathog,* 14(1), p. 46.

Sarosiek, I., Schicho, R., Blandon, P. & Bashashati, M., 2015. Urinary metabolites as noninvasive biomarkers of gastrointestinal diseases: A clinical review. *World journal of gastrointestinal oncology,* 8(5), pp. 459-465.

Vila, A. V. et al., 2023. Faecal metabolome and its determinants in inflammatory bowel disease. *Gut,* 72(8), pp. 1472-1485.