Metabolomic Analysis in Pediatric Crohn´s Disease Patients: A Machine Learning Approach

Ricardo Suarez Suarez
University of Alberta
suarez@ualberta.ca

Namitha Guruprasad
University of Alberta
ngurupra@ualberta.ca

Ganesh Tata
University of Alberta
gtata@ualberta.ca

Santiago Miro Trejo
University of Alberta
mirotrej@ualberta.ca

Nabil Al Asmer
University of Alberta
alasmer@ualberta.ca

Abstract

Inflammatory Bowel Diseases (IBD) are disorders in the intestinal tract with no permanent cure and can often lead to hospitalizations, surgeries, and life-threatening complications. Furthermore, IBD are more severe in children than in adults and carry a risk of growth failure, delayed puberty, depression, anxiety, and cancer.

Metabolomics allows us to analyse the metabolic changes in the gut microbial metabolism by identifying and quantifying the compounds in biological samples. Therefore, metabolomics provide a unique opportunity to associate pediatric Crohn's disease (pCD) with an altered multivariate metabolite profile.

The objective of this work is to perform an exploratory data analysis on serum metabolomics from 56 pCD patients. Samples were collected twice: once, as a study initiation (Visit One) and second, for completion (Visit Four). Then, the collected serum underwent a liquid chromatography followed by a mass spectroscopy process to identify 130 compounds. Serum metabolites are analysed by applying Unsupervised (U.ML) and Supervised (S.ML) Machine Learning algorithms. U.ML is used to reduce the dimensionality of the data and identify clusters, whereas S.ML is used to develop regression and classification algorithms to assess the correlation between metabolites and disease activity.

Results show that Tryptophan is the metabolite that ranks the highest in the feature importance scoring. Histidine is another metabolite that correlates well with severity of the disease. In regression analysis, it was seen that Partial Least Squares (PLS) with Linear Optimal Low-rank (LOL) reduction setting incurred the least Mean Squared Error (MSE) of 0.8806. In classification process, Random Forest with Linear Discriminant Analysis (LDA) setting achieved the least False Negative Rate of 0.53 and K-Nearest Neighbours (KNN) with Principal Component Analysis (PCA) setting yielded the highest accuracy score of 63%.

Thus, we inferred that the above alternatives did not have any exceptional impact on our findings. The major reason for this might be the scarcity of data and imbalance or under-representation of some data labels.

1 Introduction

The two main subtypes of IBD are Crohn disease (CD) and ulcerative colitis (UC). The CD inflammation occurs anywhere along the gastrointestinal tract with skip lesions, while the UC inflammation is restricted to the colon in a contiguous fashion [1].

Submitted to STAT 541 Final Project. Do not distribute.

The worldwide incidence of IBD is soaring, especially in some Western countries, such as Canada, where the rates are projected to be as high as 1% of its entire population by 2030 [2]. Moreover, the prevalence in children has been significantly rising, with up to 25% of diagnoses occurring in the pediatric population [3].

IBD in children is related with a more severe disease course that in adults [4]. Growth and development concerns are unique to pediatric IBD (pIBD). Moreover, younger age of diagnosis in pIBD relates to a more complicated condition where patients present poorer outcomes and an increased need for surgery [5]. Also, diagnosis with IBD in childhood has been associated with increased risk of cancer and mortality in adulthood [6].

Current findings show that genetic susceptibility, diet changes, and environmental factors might be the factors connected with the disease development. It is an alarming concern that the development of IBD remains poorly understood in the community and needs to be explored for further analysis. Some novel procedures can be used to further study the aetiology of IBD which include the metabolomics, metagenomics, 16s-RNA analysis of gut microbiome, and epigenetic analysis.

Metabolites are compounds that are necessary for or formed during metabolism which comprises both host and microbe factors that reflect the real-time microenvironment of the gut. Metabolomics quantify the host and microbe metabolites to distinguish between CD, UC, and healthy controls, but there has been limited investigation done in paediatric IBD metabolomics [7]. This provides the research community an excellent opportunity to examine metabolomic profiles and their connection with disease severity among pCD patients.

2 Methodology

2.1 Data Acquisition

ImageKids is a multicentre, prospective cohort observational study that was conducted to develop magnetic resonance enterography (MRE) indices for pCD. The study is conducted over 18 months with serum and urine specimens collected as a study initiation (Visit One or V1) and completion (Visit Four or V4) processes for 56 pCD patients.

Collected serum samples are sent to University of Alberta, and then to The Metabolomics Innovation Centre (TMIC) for metabolomics analysis. Serum undergoes a liquid chromatography followed by mass spectroscopy procedure as a result of which 130 compounds are identified.

Serum metabolites are analyzed by applying U.ML and S.ML techniques based on Scikit-learn library [8] in Python.

2.2 Unsupervised Learning

Unsupervised Learning techniques are used to extract useful patterns from a dataset in the absence of any label information. We apply techniques such as, clustering, on the patients dataset. Since the number of features (130) is much larger than the number of samples, we also apply well-known dimensionality reduction methods and evaluate their performance using well-defined clustering metrics.

2.2.1 Dimensionality Reduction

Dimensionality reduction involves transformation of high-dimensional data to a low-dimensional data such that some meaningful information present in the original data is retained in the low-dimensional space. We use the following dimensionality reduction techniques for unsupervised learning -

- **PCA** - It is a popular approach for deriving a low-dimensional set of features from a large set of variables that finds a sequence of linear combinations of the variables exhibiting maximal variance but mutually uncorrelated. It allows us to summarize a dataset with a smaller number of representative variables that collectively explain most of the variability in the original dataset.The first principal component direction of the data is the direction along which the observations vary the most. PCA acts as a suitable technique for analysing our metabolomic data [9].

- **UMAP** [10] - Uniform Manifold Approximation and Projection technique transforms the data by constructing a neighbourhood graph in the original data space and obtaining a low-dimensional space that preserves the clusters in the original data space.
- **MDS** - Multi-dimensional scaling is a non-linear technique for embedding data in a low-dimensional space which maps points residing in a high-dimensional space to a low-dimensional space while preserving the distances between those points as much as possible. Because of this, the pairwise distances between points in the lower-dimensional space are matched closely to their actual distances.
- **PCA + UMAP** - We use PCA to obtain the first k principal components which account for the most variation in the data. Then, we use UMAP to map the data to the 2-dimensional space. Applying PCA allows us to get some computational speedup for UMAP. This also helps in reducing the noise in the data without severely distorting the distance between the data points.
- **MDS + UMAP** - We use MDS for embedding the dataset in a low-dimensional space, which preserves the distances between the points. Then, we use UMAP to map the new dataset to a 2-dimensional space. Using UMAP directly leads to error that arises from changes in the global structure of the data. Applying MDS before UMAP provides an effective way to minimize this error.

Among the above mentioned techniques, PCA and UMAP provide a way to apply an inverse transformation to get high-dimensional data from low-dimensional data. Such a transformation is not possible using MDS.

2.2.2 Clustering

Clustering is an unsupervised learning technique which is used to group similar data points using a distance metric. Clustering techniques can be broadly classified into density-based clustering, centroid-based clustering and hierarchical clustering. Since density-based clustering techniques require a large amount of data, we cannot apply them to our dataset. Both hierarchical clustering and centroid-based clustering yield similar results on a smaller dataset. Hence, we perform our experiments using the K-Means clustering algorithm that groups our data points into k clusters by assigning a cluster label to each point based on its distance from the mean of each cluster. The process is repeated until a convergence criterion is satisfied.

2.2.3 Partial Least Squares

Partial least squares-discriminant analysis (PLS-DA) has been extensively used in metabolomics[11]. PLS is a multivariate regression method where information from a data space of a larger number of variables is projected into that with a smaller number of variables[12].

In this work, the associated clusters are analyzed using PLS-DA, to identify metabolites most responsible for cluster definition. Consequently, the variables importance in projection (VIPs) corresponding to the geography of the patients are presented below 2

PLS is also used to maximize the co-variance between the metabolites and the severity of the disease. Therefore, PLS is explained with more detail in the S.ML section.

2.3 Supervised Learning

One of the most important subcategories of Machine learning is Supervised learning which trains various algorithms to predict the outcomes or classify the data. Usually, the dataset (X, Y) comprises a set of features (X) that are associated with a target response (Y). Applying this concept to the metabolomics dataset helps in deducing a set of conclusions at scale, such as finding the metabolites that have a major impact on the severity of Crohn's disease. Figure 1 shows a fundamental set of actions for evaluating a bunch of models based on the key features extracted and training of the models.

S.ML is used to assess correlation between metabolites and disease activity. Therefore, PICMI index is used to label and classify patients in three different categories: remission, moderate-mild, and

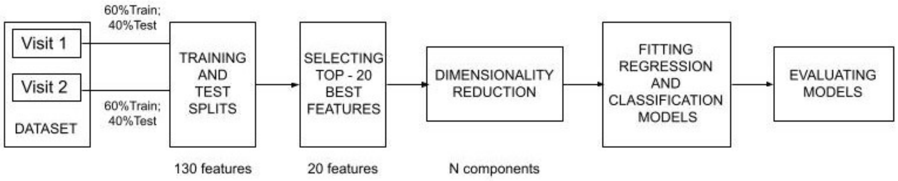


Figure 1: Set of action for evaluating models.

severe disease. PICMI is a valid, reliable, and responsive index for assessing transmural inflammation in pCD[13]. The class labels in our dataset is created using the PICMI scores as follows:

- < 40 PICMI score is considered on Remission.
- $41 - 70$ is considered Mild+Moderate.
- > 70 is considered Severe.

2.3.1 Pre-Processing of Data

Splitting the Data

The Metabolomics dataset is composed of 56 patients' records from two different visits (Visit 1 and Visit 2) that are at an interval of 4 months. Due to the brevity of dataset, the "Mild+Moderate" class label is underrepresented and emanates the problem of its under-representation. To combat this issue, we extract equal proportions (60%) of every class label from both the visits for training and the remaining for testing procedures.

```
Training : class labels count for 1st visit
Counter({0: 17, 2: 12, 1: 4})
Test : class labels count for 1st visit
Counter({0: 11, 2: 9, 1: 3})

Training : class labels count for 2nd visit
Counter({0: 21, 2: 8, 1: 4})
Test : class labels count for 2nd visit
Counter({0: 15, 2: 6, 1: 2})
```

Scaling of Data

Once the unified set is obtained, the data is scaled by removing the mean and scaling to unit variance. This is known as Standard Scaling. The scaling process avoids the poor behavior of the features from the data that often occurs when the data is not following a consistent distribution.

Selecting Top 20 Best Features

Originally, the dataset is constituted of 130 metabolomic features which are substantially more than the number of samples per visit. Since the sample size is typically orders of magnitude smaller than the dimensionality of the data, this results in the curse of dimensionality. Hence, generating a low-dimensional representation for valid inferences and preserving the discriminating information (for instance, does the patient suffer from Crohn's disease?) is essential. There is a need for interpretable supervised dimensionality reduction methods that scale to a wide data with billions of dimensions with compelling statistical guarantees.

Initially, we applied a simple baseline approach to select the best representative features based on univariate statistical tests. This acted as a pre-processing step to fit any estimator with scaled data. The sklearn's *SelectKBest* wrapper proved useful to extract all highest scoring features in the dataset. We used the following scoring functions for our dataset:

- *fregression*: This function is utilized for a calculating cross-correlation between every sample (X) and the target response (Y) as follows:

$$crosscorrelation = \frac{(X[:, i] - mean(X[:, i])) * (Y - mean(y))}{std(X[:, i] * std(y)} \tag{1}$$

- *fclassif*: This function is utilized for calculating ANOVA value for the provided samples and target response.

These scores generated from the functions are converted into univariate F-scores and then to p-value statistics, thus identifying potentially predictive features for an estimator irrespective of the sign of the association of the target variable.

In addition to this, we applied a few classical statistical starter approaches to the top key features extracted ($p >> n$) along with a novel method called LOL that adapts to the complexity of the data, is robust to outliers, and is computationally efficient. The strategies pursued in our cases is as follows:

- LDA: We explored the application of linear discriminant analysis (LDA) to the Metabolomics features because this technique reduces dimensions while preserving the global structure of data. We drew our inspiration from [14],to perform scaling of every Metabolite such that the distribution of data in each class ("Remission", "Mild+Moderate", "Severe") is as close as possible to a standard normal distribution in a transformed space. Distances between the objects in this space are hence made meaningful by estimating the centroids and their distances, since the space is homogeneously deformed, and length along each dimension is expressed in multiples of intra-class variation.
- PCA: We analyzed the conventional Principal Component Analysis (PCA) to show the impact of sample reduction on the process of reducing dimensions for supervised learning [15] . We took into account the class information for calculating the class-conditional probabilities and optimizing the ratio of between-class variance to the within-class variance of the transformed data by calculating the covariance matrix. The computation of the components is based on the eigenvalue decomposition of the covariance matrix S as follows:

$$w \leftarrow eigenDecomposition(S = \sum_{i=1}^{n} (x_i - mean(X))(x_i - mean(X))^T) \tag{2}$$

- NCA: We considered the Neighbourhood Component Analysis (NCA) to carry out the supervised dimensionality reduction [16] where the Metabolites are projected onto a linear subspace consisting of the directions which minimize the NCA objective.. This distance metric learning algorithm helped us to maximize the stochastic variant of the leave-one-out k-nearest neighbors (KNN) score on the training set. Optimal linear transformation maximizes the sum over all the patient samples and records the probability of the correctly classified samples on the basis of a stochastic nearest neighbors rule:

$$P_{ij} = \frac{exp(-||Lx_i - Lx_j||^2)}{\sum_{k \neq i}^{n} exp(-||Lx_i - Lx_j||^2)}, P_{ii} = 0, \tag{3}$$

  where, p_{ij} is the softmax over Euclidean distances in the learned embedded space.
- LOL: Finally, we performed the Linear Optimal Low-rank projection (LOLp) that aimed to integrate the class-conditional means [17] to substantiate classification of improved representations of the Metabolites with good computational efficiency and scalability. Our inspiration to use this technique stems from its exceptional performance on genomics datasets with more than thousands of features when compared to other scalable linear linear dimensionality reduction techniques in terms of accuracy.

2.3.2 Regression Techniques

Regression analysis is a combination of statistical techniques that aims to establish relationship between independent and dependent variables along with minimizing the differences between the regression estimate or prediction and the true value. This helped us to analyze the dependence of the Metabolites on the PICMI scores and eventually on the severity of the Crohn's disease.

Our basic aim is to train our model to classify into three classes. A set of bins regarding the PICMI cut offs are introduced in the data to transform into the classes needed for the following regression techniques:

- Linear Regression: As a baseline technique, we performed simple linear regression for modelling the Metabolomics features on the severity of the disease keeping in mind the minimization of error rates associated with the predicted and the true values.
- PLS: We incorporated Partial Least Squares (PL2S) as a technique to maximize the covariance between the Metabolites and the set of target responses (severity of the disease). This helped on emphasizing on the predictive modelling which reduced a number of factors and worked well for highly collinear/noisy data which often occurs in the metabolomics experiments [18].
- Ordinal Regression: Since our dataset belongs to the field of metabolomics, usually the variables are in natural ordering format (ordinal scale) [19]. Ordinal regression acted as a fence between Regression and Classification techniques. We used it as an inclination towards regression to calculate the cumulative probabilities of the Severity scale (response variable) as a function of the Metabolites using the logistic function for constructing a Ordinal Logit model.

2.3.3 Classification Techniques

We also applied classification techniques for the analysis of our data. The current objective is to predict the outcome of the data set as a qualitative output. This is an intuitive step in our analysis, since depending on the PICMI score each patient has, it is considered at a different stage on the severity of their disease.

When talking about a classification technique, if our predictor G(x) takes values in a discrete set G, we can always divide the input space into a collection of regions labeled according to the classification. The boundaries of these regions can be rough or smooth, depending on the prediction function. There are several different ways in which linear decision boundaries can be found. In this paper we took four classification methods such as:

- Logistic Regression: Model the posterior probabilities of the K classes via linear functions in x, while ensuring that they sum to one between [0,1]. This model has the form:

  $$log\frac{Pr(G=K-1|X=x)}{Pr(G=K|X=x)} = \beta_{(K-1)0} + \beta_{K-1}^{T}x. \tag{4}$$

  We decided to start our classification with this method, since this is arguably the most simple algorithm to work on [20]. The results obtained from this is used as a control reference for the following methods. Moreover, logistic regression is very simple to interpret, which would be helpful for future analysis.
- Support Vector Machine: Taking things one step further, the decision of using Support Vector Machine (SVM) is a natural transition. So far we have found linear boundaries in the input feature space. However, here it is possible to to make the procedure more flexible by enlarging the feature space using polynomials and splines. Linear boundaries in the enlarged space achieve better training-class separation, and translate to nonlinear boundaries in the original space. Furthermore, the dimension of the enlarged space is allowed to get very large, and even to infinity. This is a more complex technique for classification, but at the same time it is more flexible for more entangled data. Moreover, one strong motivation to use SVM is that it has shown in previous works, that the method works well in the study of metabolomics, which greatly justified our decision. [18]
- Random Forest Classifier: The essential idea in Random Forest is to average many noisy unbiased models, and hence reduce the variance. Trees are ideal candidates for bagging, since they can capture complex interaction structures in the data, and if grown sufficiently deep, have relatively low bias. Since trees are notoriously noisy, they benefit greatly from the averaging. When used for classification, a random forest obtains a class vote from each tree, and then classifies using majority vote. This method was also considered for a classification analysis. It displayed a different behavior than that of SVM, leading to a more complete

inspection of our data. Furthermore, RF was also an accepted method from previous work in the study of metabolomics, the same way as SVM.[18] These are clear indicators that our methodology is coherent.

- K-nearest Neighbor (KNN): KNN is a non-parametric method for performing classification. The decision for selecting this method was made by mere curiosity of looking at the results provided by it. Since this classification is memory-based, and requires no model to be fit [20], it is not expected for it to be as precise as the previous two. Nevertheless, there is always the possibility of finding interesting patterns in other models.

3 Experiments and Results

Table 1: Clustering results for V4 data

Dimensionality Reduction	**Silhouette Coefficient**	**Calinski-Harabasz Index**	**Reconstruction Error**
—	0.289	21.331	—
PCA	0.661	105.429	**0.058**
MDS	0.471	47.890	—
UMAP	0.781	454.372	45.919
PCA + UMAP	0.947	**7517.397**	183.468
MDS + UMAP	**0.948**	7512.461	—

Table 2: Top VIP Serum Metabolites distinguishing Geographical groups

Metabolite	**VIP Score**
LYSOC20:3	6.06
Sarcosine	3.18
Spermidine	3.03
C5	2.24
Creatinine	1.95
Spermine	1.93
Asymmetric dimethylarginine	1.89
Threonine	1.76
Putrescine	1.69
C18:1	1.57

3.1 K-Means, Dimensionality Reduction and PLS-DA

As indicated in previous metabolomics research [21], different scaling and transformation methods have their own advantages and disadvantages. We applied pareto scaling and log transformation to our dataset since it yielded the best clusters when K-Means is performed on the raw features (metabolites). The specific parameters chosen for the different techniques are as follows (all dimensionality reduction techniques result in a 2-dimensional embedding) -

- **KMeans** - 2 clusters are specified. Euclidean distance is used to form the clusters.
- **UMAP** - The number of neighbours is chosen to be 20. The distance metric that lead to the best results (i.e. well-formed clusters) is cosine similarity.
- **PCA + UMAP** - We performed dimensionality reduction from 130 features to k features using PCA. The k id determined by performing clustering with PCA + UMAP for different values of k. Using the clustering evaluation metrics, k id chosen to be 24.
- **MDS + UMAP** - An analysis similar to **PCA + UMAP** is performed to determine the k value as 35.

3.1.1 Evaluation Methodology

We use the Silhouette Coefficient [22] [8] and the Calinski-Harabasz Index [23] to evaluate the clustering results since these metrics do not require label information for the data points. The Silhouette Coefficient measures how close each data point is to its own cluster when compared with other clusters. The coefficient's value is between -1 and 1, and a higher value indicates better clusters. The Calinski-Harabasz Index measures the ratio between within-class disperstion and between-class dispersion, where a higher value indicates well-formed clusters.
In addition to evaluation metrics for clustering, we also evaluated the dimensionality reduction techniques using the reconstruction error metric. Performing dimensionality reduction can lead to information loss and one way to quantify this loss is to determine if the original data can be reconstructed from the low-dimensional space. Hence, for PCA and UMAP, we applied their inverse transform on the low-dimensional data to obtain the reconstructed features, and then computed its MSE with the original data to obtain the reconstruction error. A lower error showed that the original metabolite information for each patient can be retrieved even from a low-dimensional space. Since MDS does not have an inverse transform, we did not report the reconstruction error for it.

3.1.2 Clustering results

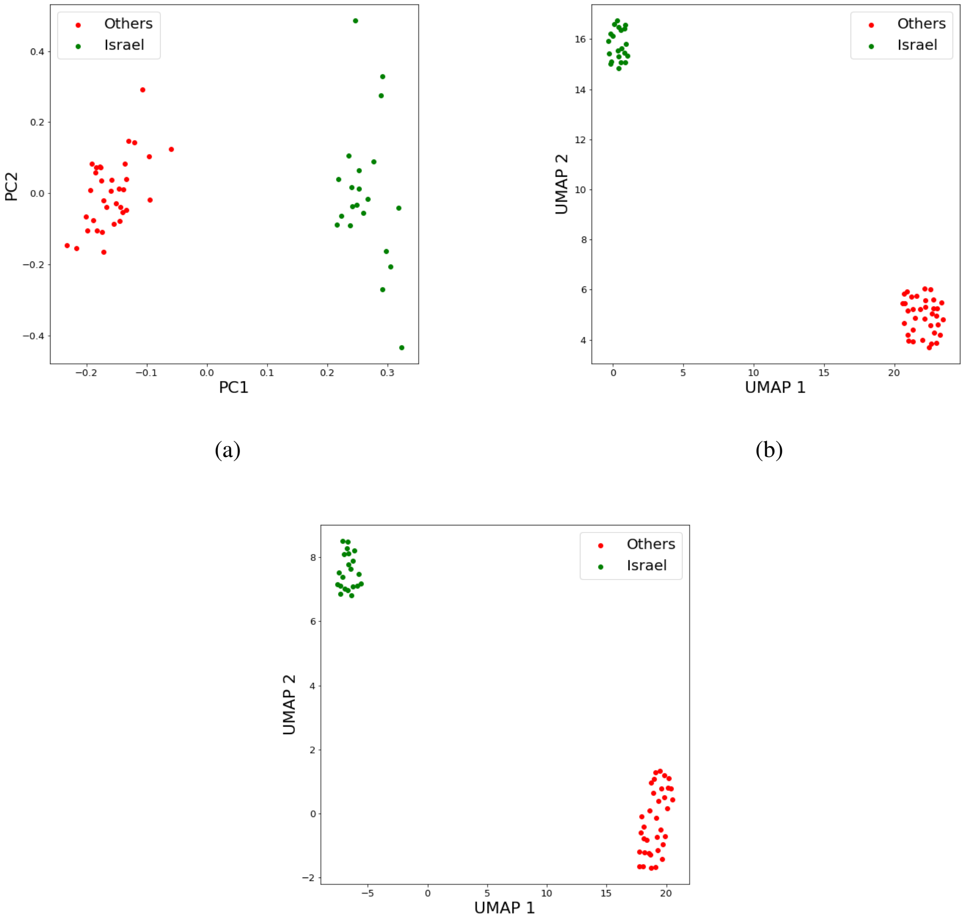
Table 1 depicts the results for applying K-means with different dimensionality reduction techniques. Through the clustering evaluation metrics, it can be observed that K-means with MDS + UMAP and K-means with PCA + UMAP yielded the best clusters. However, the reconstruction error is lower when only PCA is used. Hence, we want to highlight that both methods have their pros and cons. The formation of better clusters comes at the cost of losing some information from the original data. It is also observed that the first two component obtained through PCA explain around 44% of the variance in the data.
Figure 2 shows the cluster visualization for K-means with PCA, PCA + UMAP and MDS + UMAP. The clusters clearly separate patients from Israel and patients from other countries. Similar results are observed for other dimensionality reduction techniques as well.

Therefore, geography was found as an important environmental factor driving distinct patterns of the metabolomics profile.

3.2 Feature Selection

Models thrive under constructive criticism that gives us an opportunity to make enhancements, corrections and improvements for better performance. Hence, there are plans of actions, strategies and protocols with different metrics for analysing the behaviour of the models. It is mandatory to consider the model pipeline, constraints and environments of operation. For fair comparison of models, existing evaluation metrics are used to measure the regression and classification techniques. Our dataset consists of 130 features for analysis. Upon applying sklearn's *SelectKBest* module, we were able to extract top 20 best features along with their feature scores as follows:

```
               Feat_names   F_Scores
               Tryptophan   12.001081
              Citric acid   8.440411
                   16:1SM   6.037837
                     C5DC   5.986706
                    C10:1   5.763595
                    C5MDC   5.632694
                  PC32:2AA  5.380965
                 Tyrosine   5.277203
                Histidine   5.074206
    Trimethylamine N-oxide  3.865438
  beta-Hydroxybutyric acid  3.818782
               Methionine   3.734028
                   Lysine   3.652867
                   18:1SM   3.480790
                 Creatine   3.335337
                LYSOC18:2   3.225121
```



(a)

(b)

(c)

Figure 2: Clustering visualization using K-means with (a) PCA (b) PCA + UMAP (c) MDS + UMAP

```
                      C7DC    3.192890
5-Hydroxyindoleacetic acid    2.669419
                      C4OH    2.527650
                  PC40:1AA    2.465284
```

3.3 Regression

Error estimators calculate the average squared difference between the predicted values of the target response by the regression models and the true/actual values which correspond to the expected values of the loss associated with the predicted values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i^{'})^2, \tag{5}$$

where Y_i is true value and $Y_i^{'}$ is the predicted value of the regression model; $(Y_i - Y_i^{'})^2$ is the square of errors upon which an average operation is performed to generate the Mean Squared Error which is sample dependent. Following is the Table 3 for MSE for different regression models.

Table 3: Regression results for data

Regression Model	Dimension Reduction Method	Mean Square Error MSE	Root Mean Squared Error RMSE
Linear Regression			
	LDA	1.1366	1.0661
	PCA	0.8811	0.9387
	NCA	1.0355	1.0176
	LOL	0.9113	0.9547
PLS			
	LDA	1.1366	1.0661
	PCA	0.8807	0.9384
	NCA	0.9855	0.9927
	LOL	0.8806	0.9384
Ordinal			
	LDA	1.5	1.2247
	PCA	1.4130	1.1887
	NCA	1.3260	1.1516
	LOL	1.5	1.2247

3.4 Classification

- Confusion Matrix: An array is constructed for salient mapping mechanism to perform a thorough breakdown of true and false class labels while judging a model, mainly for classification. Here, in Table 4 a confusion matrix is used to gauge our model's ability to classify the patient records to the severity scale.

	PREDICTED		
	Positive	Negative	
ACTUAL	Positive	True Positive (PP)	False Negative (PN)
	Negative	False Positive (NP)	True Negative (NN)

Table 4: Confusion matrix

- Accuracy: It is the most salient metric for evaluation a model's performance which calculates the percentage / proportion of the correctly classified instances by the model. Here, accuracy is computed to estimate the number of correctly classified patient records to their respective labels

$$Accuracy = \frac{(PP + NN)}{(PP + NN + PN + NP)} \tag{6}$$

  Table 5 highlights accuracy of all models.

- False Negative Rate of the "Severe" class - False Negative rates are very dangerous in medical data analysis as it has the tendency to incorrectly classify a patient to "Remission" or "Mild+Moderate" instead of "Severe" which can lead to fatality.

Table 5: Accuracy and False Negative Rate

Classifying Model	Dimension Reduction	Accuracy	False Negative Rate
Logistic Regression			
	LDA	59%	0.6
	PCA	50%	0.73
	NCA	50%	0.73
	LOL	48%	0.73
SVM			
	LDA	50%	0.8
	PCA	41%	0.93
	NCA	39%	0.93
	LOL	37%	1
Random Forest			
	LDA	59%	0.53
	PCA	52%	0.66
	NCA	57%	0.66
	LOL	46%	0.8
K-nearest Neighbors			
	LDA	54%	0.73
	PCA	63%	0.6
	NCA	61%	0.6
	LOL	50%	0.8

4 Discussion and Conclusion

From the cluster analysis, V4 clustering is mainly attributed to polyamines involved in regulation of cell growth/death and inflammation. Furthermore, demographics is found as an important environmental factor driving distinct patterns of the metabolomics profile.

From the feature selection step, we observed that Tryptophan ranks highest in the feature importance scoring.In fact, Tryptophan has been previously identified as significantly altered in the blood of IBD patients compared to controls [24]. Histidine is another metabolite that seems to correlate well with disease severity. As a matter of fact, histidine is known to be involved in the mediation of oxidative stress, potentially influencing intestinal inflammation[25]. Finally, lysisne and tyrosine are metabolites that have also been shown to be particularly discriminatory between IBD and control groups[26].

In the unsupervised learning approach, we applied the K-Means clustering algorithm and observed two clusters, Israel and non- Israel. Further, we used some of the dimensionality reduction and we observed that applying MDS or PCA then UMAP is more efficient than the individual techniques. In addition to this, we evaluated the results by computing some of the evaluating measurements, the Silhouette Coefficient [22] [8], the Calinski-Harabasz Index [23], and the reconstruction error.

In the supervised learning approach, we observed that the number of metabolites is reduced from 130 to 20 best features based on the feature scores generated by sklearn's SelectKBest wrapper module. In regression analysis, it is seen that PLS with LOL reduction setting incurred the least MSE of 0.8806. In classification process, Random Forest with LDA setting achieved the least False Negative rate of 0.53 and KNN with PCA setting yielded the highest accuracy score of 63Before carrying out the experiments on the subset of features, we performed regression and classification techniques on the entire raw feature set i.e. on all the 130 features. Upon performing this experiment, we noticed that the performance metrics and error rates remained almost constant. This suggests that it is a wiser approach to reduce the large Metabolites feature set into a smaller set of essential features that would eventually ease the process of carrying out regression/classification tasks with a slightly better performance. In addition to this, we also performed L1 regularization on our dataset during the regression analysis. The reason behind the choice of L1 regularizer is size of our Metabolites feature set and applying this methodology would provide sparse solutions. With this motive, we applied the L1 penalty on our regressors but did not find any remarkable differences in our results. Considering the attempts to make our models better, we inferred that the above alternatives did not have any exceptional impact on our findings. The major reason for this might be the scarcity of data and imbalance of some data labels like "Mild+Moderate".

5 Future Work

The major thing that we would yearn for in the future is to quench our thirst of not having sufficient data. We aim to perform our experiments on a larger dataset with classes that are proportionally represented by the samples. Having more data might support our findings in a better way and deduce more relationships between the Metabolites and the demographics/severity scale. A boon to us would be to obtain data from several visits for performing time-series analysis. This would provide us an insight into the specific set of Metabolites that contribute to the change in the status of the patient from more severe to less severe. This might also help us to study the trend of Crohn's disease in different geographical locations and impact of their diet on the severity of the disease. On statistics front, we would like to concentrate on the tuning of Hyperparamters for a better performance of our regression/classification models by adopting the automatic procedure of detecting the optimal hyperparameters in terms of optimization strategies such as typical brute force method like Grid search or Random search. For instance, hyperparameters for random forest could be:

- The number of trees built.
- “gini”/”entropy” measures for checking the quality of the splits and purity of the nodes.
- Control the depth of the tree and other pruning techniques

References

[1] R Balfour Sartor. Mechanisms of disease: pathogenesis of crohn's disease and ulcerative colitis. *Nature clinical practice Gastroenterology & hepatology*, 3(7):390–407, 2006.

[2] Stephanie Coward, Fiona Clement, Eric I Benchimol, Charles N Bernstein, J Antonio Avina-Zubieta, Alain Bitton, Mathew W Carroll, Glen Hazlewood, Kevan Jacobson, Susan Jelinski, et al. Past and future burden of inflammatory bowel diseases based on modeling of population-based data. *Gastroenterology*, 156(5):1345–1353, 2019.

[3] Yizhou Ye, Sudhakar Manne, William R Treem, and Dimitri Bennett. Prevalence of inflammatory bowel disease in pediatric and adult populations: recent estimates from large national databases in the united states, 2007–2016. *Inflammatory Bowel Diseases*, 26(4):619–625, 2020.

[4] Ryan Carvalho and Jeffrey S Hyams. Diagnosis and management of inflammatory bowel disease in children. In *Seminars in pediatric surgery*, volume 16, pages 164–171. Elsevier, 2007.

[5] Gwenola Vernier-Massouille, Mamadou Balde, Julia Salleron, Dominique Turck, Jean Louis Dupas, Olivier Mouterde, Véronique Merle, Jean Louis Salomez, Julien Branche, Raymond Marti, et al. Natural history of pediatric crohn's disease: a population-based cohort study. *Gastroenterology*, 135(4):1106–1113, 2008.

[6] Mikkel Malham, Christian Jakobsen, Anders Paerregaard, Lauri J Virta, Kaija-Leena Kolho, and Vibeke Wewer. The incidence of cancer and mortality in paediatric onset inflammatory bowel disease in denmark and finland during a 23-year period: a population-based study. *Alimentary pharmacology & therapeutics*, 50(1):33–39, 2019.

[7] Elizabeth A Scoville, Margaret M Allaman, Caroline T Brown, Amy K Motley, Sara N Horst, Christopher S Williams, Tatsuki Koyama, Zhiguo Zhao, Dawn W Adams, Dawn B Beaulieu, et al. Alterations in lipid, amino acid, and energy metabolism distinguish crohn's disease from ulcerative colitis and control subjects by serum metabolomic profiling. *Metabolomics*, 14(1):1–12, 2018.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[9] Gift Nyamundanda, Lorraine Brennan, and Isobel Claire Gormley. Probabilistic principal component analysis for metabolomic data. *BMC bioinformatics*, 11(1):1–11, 2010.

[10] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.

[11] E Szymanska, E Saccenti, AK Smilde, and JA Westerhuis. Double-check: validation of diagnostic statistics for pls-da models in metabolomics studies. metabolomics 8. *S3–S16*, 2012.

[12] Tahir Mehmood, Kristian Hovde Liland, Lars Snipen, and Solve Sæbø. A review of variable selection methods in partial least squares regression. *Chemometrics and intelligent laboratory systems*, 118:62–69, 2012.

[13] Greer M. Pratt L Focht G., Kuint R. Development, validation and evaluation of the pediatric inflammatory crohn's magnetic resonance enterography index (picmi) from the imagekids study. *Journal of Crohn's and Colitis*, pages 1–27, 2022.

[14] Francisco JH Heras and Gonzalo G de Polavieja. Supervised dimensionality reduction by a linear discriminant analysis on pre-trained cnn features. *arXiv preprint arXiv:2006.12127*, 2020.

[15] Mykola Pechenizkiy, Seppo Puuronen, and Alexey Tsymbal. The impact of sample reduction on pca-based feature extraction for supervised learning. In *Proceedings of the 2006 ACM symposium on Applied computing*, pages 553–558, 2006.

[16] Sam Roweis, Geoffrey Hinton, and Ruslan Salakhutdinov. Neighbourhood component analysis. *Adv. Neural Inf. Process. Syst.(NIPS)*, 17(513-520):4, 2004.

[17] Joshua T Vogelstein, Eric W Bridgeford, Minh Tang, Da Zheng, Christopher Douville, Randal Burns, and Mauro Maggioni. Supervised dimensionality reduction for big data. *Nature communications*, 12(1):1–9, 2021.

[18] Piotr S Gromski, Howbeer Muhamadali, David I Ellis, Yun Xu, Elon Correa, Michael L Turner, and Royston Goodacre. A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. *Analytica chimica acta*, 879:10–23, 2015.

[19] Alan Agresti. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2010.

[20] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[21] Robert A van den Berg, Huub CJ Hoefsloot, Johan A Westerhuis, Age K Smilde, and Mariët J van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, 7(1):1–15, 2006.

[22] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[23] Calinski T Harabasz and M Karoński. A dendrite method for cluster analysis. In *Communications in Statistics*, volume 3, pages 1–27. 1974.

[24] Kate Gallagher, Alexandra Catesson, Julian L Griffin, Elaine Holmes, and Horace RT Williams. Metabolomic analysis in inflammatory bowel disease: a systematic review. *Journal of Crohn's and Colitis*, 15(5):813–826, 2021.

[25] Kohei Sugihara, Tina L Morhardt, and Nobuhiko Kamada. The role of dietary nutrients in inflammatory bowel disease. *Frontiers in Immunology*, 9:3183, 2019.

[26] Yi Jie Weng, Huo Ye Gan, Xiang Li, Yun Huang, Zheng Chao Li, Hui Min Deng, Su Zuan Chen, Yu Zhou, Li Sheng Wang, Yan Ping Han, et al. Correlation of diet, microbiota and metabolite networks in inflammatory bowel disease. *Journal of Digestive Diseases*, 20(9):447–459, 2019.