Report on Using ML to Study Autoimmune Diseases

Prediction, Analysis and Further Study

Prepared for

Dr. Meenal Kowshik

Department of Biological Sciences

By

Laukik B Nakhwa- 2020B1A81932G

Gaurang Aswal -2020B1A71960G

Jimit Shah -2020B1A72097G

Classification of Paediatric Inflammatory Bowel Disease Using Machine Learning(Visual Approach)

Pediatric inflammatory bowel disease (PIBD), comprising **Crohn's disease (CD), ulcerative colitis (UC), and inflammatory bowel disease unclassified (IBDU),** is a group of autoimmune inflammatory conditions affecting children. We utilised a **Supervised Learning Model** based on a **Random Forest** algorithm to classify PIBD into its subtypes. (1)

A dataset with **Histological and Endoscopic** data was utilised. (https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-017-02606-2/MediaObjects/41598_2017_2606_MOESM1_ESM.xls/)

The central feature of inflammatory bowel disease is chronic gastrointestinal (GI) tract inflammation. Symptoms of PIBD include diarrhoea, abdominal pain, blood in the stool, and weight loss. Although both Crohn's disease and ulcerative colitis fall within the same disease group, **there are often differences in disease location within the bowel, observable through endoscopic and histological assessment.**

A non-continuous inflammation of the entire gastrointestinal system typically characterises Crohn's disease. In contrast, the inflammation pattern of ulcerative colitis is continuous and restricted to the colon and rectum.

Diagnosis of PIBD is challenging, the aetiology/cause of the disease is not fully understood, and deciding on management and prognostication is complex. **The accuracy of diagnosis in PIBD is critical to prompt and effective treatment. The treatment for**

PIBD is highly dependent on disease location and extent and accurately classifying it as CD, UC, and IBDU.

As per the dataset, Endoscopic and histological data were collected for 287 patients: 178 patients with Crohn's disease, 80 with ulcerative colitis, and 29 with inflammatory bowel disease. Machine learning was applied to 239 patients (CD=143, UC=97, IBDU=29). Females account for 37% (107) of the individuals in the dataset. The average age of onset was 11.5 years (range 1.6 to 17.6 years). Twenty-six (9%) of patients were diagnosed below six years of age (very-early onset IBD). The remaining 48 patients (CD=35, UC=13, average age of onset 13.2 years) were used to validate the model.

Logic and Scoring: Ten gastrointestinal (GI) locations were investigated for the presence of macroscopic and microscopic abnormalities: mouth, oesophagus, stomach, duodenum, ileum, ascending colon, transverse colon, descending colon, rectum, and perianal. Depending on tissue abnormalities, clinical observations were converted into numerical variables [−1, 0, +1]. At each location, abnormal tissue observations were coded as +1 and normal as −1. Null values (0) were assigned for missing data, such as in the case of restriction at endoscopy.

Results :

Our current ML model achieved a **65%** accuracy in predicting Inflammatory Bowel Disease (IBD) subtypes based on body location-specific input values.

Optimisation Plan:

To improve accuracy, we can implement SVMs and focus on:

Data Preprocessing: Addressing missing values and scaling features.

Exploring feature engineering.

Hyperparameter Tuning: Systematic search for optimal kernel and regularisation parameters.

Ensemble Methods: Evaluating bagging or boosting to enhance model generalisation.

Class Imbalance Handling: Mitigating class imbalance through oversampling or undersampling.

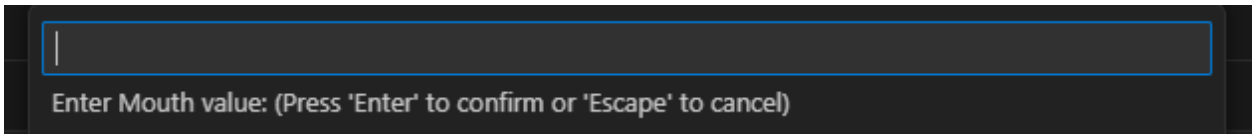
Feature Importance Analysis: Identifying and refining critical features.

Regularisation: Tuning regularisation parameters for better generalisation.

Diverse Metrics: Assessing precision, recall, F1 score, and AUC-ROC for comprehensive evaluation.

These optimisations aim to surpass the initial 65% accuracy, ensuring a more robust and reliable IBD subtype prediction model.

```
Test Accuracy: 0.6458333333333334

Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.71      0.77        35
           1       0.00      1.00      0.00         0
           2       0.60      0.46      0.52        13

    accuracy                           0.65        48
   macro avg       0.48      0.73      0.43        48
weighted avg       0.77      0.65      0.70        48


Confusion Matrix:
 [[25  6  4]
 [ 0  0  0]
 [ 5  2  6]]
```



Taking user input

```
...   Predicted Diagnosis: Ulcerative Colitis
```

Prediction based on user input by studying the training and test dataset.

Code

Test Accuracy: This measures how accurately your model predicts the target variable on the testing data. It is the ratio of correctly predicted samples to the total samples in the testing dataset. Higher accuracy values indicate better model performance.
Classification Report: This report provides detailed metrics for each class in your dataset. For each category, it includes the following:
Precision: Precision measures how many predicted positive instances for a class were positive. It is the ratio of true positives to the sum of true and false positives.
Recall: Recall (or valid positive rate) measures how many actual positive instances for a class were correctly predicted. It is the ratio of true positives to the sum of true positives and false negatives.
F1-Score: The F1-score is the harmonic mean of precision and recall. It balances precision and memory and is often used when dealing with imbalanced datasets.
Support: The support is the number of samples in the testing dataset that belong to each class.
Confusion Matrix: The confusion matrix is a table summarising the model's performance by showing the number of true positives, true negatives, false positives, and false negatives. It provides a more detailed breakdown of the model's errors and correct predictions for each class.

Further, we looked to find out how this study of the USA can be implemented or valuable for the **Indian population**. Upon further research, we came across a study(2) that conducted a survey using centres with ISPGHAN (Indian Society of Pediatric Gastroenterology, Hepatology and Nutrition) members. These nine centres, all of which are large pediatric tertiary care referral hospitals (three public and six private), agreed to participate in the study.

Methods Data of children (≤18 years) with PIBD were collected using a proforma containing details of demographics, clinical profile, extraintestinal manifestations (EIM), investigations, disease extent, and treatment.

Results Three hundred twenty-five children [Crohn's disease: 65.2%, ulcerative colitis: 28.0%, IBD unclassified (IBDU): 6.7%, median age at diagnosis: 11 (interquartile range 6.3) years] were enrolled. 6.9% of children had a family history of IBD. Pancolitis (E4) was predominant in ulcerative colitis (57.8%) and ileocolonic (L3, 55.7%) in Crohn's disease. The perianal illness was present in 10.9% and growth failure in 20.9% of Crohn's disease cases. Steroids were the initial therapy in 84.2%, 5-amino salicylic acid in 67.3%, and exclusive enteral nutrition (EEN) in 1.3% of cases.

Conclusion: The disease location and phenotype of PIBD in Indian children are similar to the children from the West.
Therefore, we can use data from foreign countries for our research and not necessarily stick to the Indian dataset.

For contacting Indian Centres to procure data :

1. Sanjay Gandhi Postgraduate Institute Medical Sciences,
2. Lucknow, Uttar Pradesh,
3. Kanchikamakoti Child Trust Hospital, Chennai,
4. Jawaharlal Institute of Postgraduate Medical Education Research, Pondicherry,
5. All India Institute of Medical Sciences, Rishikesh,
6. PVS Memorial Hospital,
7. Aster MediCity, Kochi,
8. AMRI Hospitals, Kolkata,
9. All India Institute of Medical Sciences, Delhi,
10. Deenanath Mangeshkar Hospital & Research Centre, Pune and
11. Apollo Hospital, Delhi, India.

Using Machine Learning to Study Autoimmune Diseases via Genetic approach

Project Overview: Genomic Analysis for Disease Diagnosis

- Data Acquisition:
    - We initiated our study by collecting patient data through web scraping, encompassing both healthy individuals and those diagnosed with the target disease. The dataset included genetic information, albeit with some missing non-genetic factors like gender and age.
- Data Processing:
    - To refine our dataset, we eliminated non-genetic factors deemed irrelevant to our study, including gender and age, where data was incomplete. Subsequently, we transposed the genetic factors in Excel to enhance data clarity and facilitate downstream analyses.
- Data Labeling:
    - We introduced an additional column categorizing patients into "Healthy" and "Unhealthy" groups, reflecting their respective health statuses.
- Genetic Model:
    - Employing statistical methods, we computed p and corrected p values for each gene, offering insights into their relevance to the target disease. This allowed us to generate a sorted list of genes, prioritizing those most strongly associated with the disease.
- Diagnosis Model:
    - Building on the sorted gene data, we developed a robust diagnosis model. Through supervised machine learning, the model was trained to detect the presence of the disease based on genetic information. This model serves as a powerful tool for disease prediction and aids in understanding the genetic underpinnings of the condition.
- Significance and Implications:
    - Our approach leverages advanced data processing and modeling techniques, providing a systematic framework for understanding the genetic basis of the disease. The sorted gene list and the diagnosis model offer valuable insights for both research and clinical applications, facilitating early detection and potentially contributing to the development of targeted interventions for improved patient outcomes. Further refinements and validations will enhance the reliability and applicability of our findings.

Calculate the p-values for different genes. The p-value method is commonly used in statistical analysis to assess the significance of observed results. It is a valuable tool for identifying genes associated with particular diseases in gene expression studies. The p-value allows for comparing the strength of evidence for association across different genes, helping researchers prioritise those with the most significant relationships to the disease under investigation.

```
# Extract the gene names and initialize a DataFrame to store p-values
genes = ssc_data.columns[2:]  # Assuming gene columns start from the third column

# Create a dictionary to store gene names and their corresponding p-values
gene_p_values = {'Gene': [], 'P-Value': []}

# Perform t-tests for each gene and store the results in the dictionary
for gene in genes:
    p_value = ttest_ind(healthy_data[gene], unhealthy_data[gene]).pvalue
    gene_p_values['Gene'].append(gene)
    gene_p_values['P-Value'].append(p_value)

# Create a DataFrame from the dictionary
p_values_data = pd.DataFrame(gene_p_values)
```

Now, find the significant genes by sorting the list of corrected p-values increasingly.

We have built a Random Forest Classifier model that can predict if a person has an autoimmune disease given the gene dataset with an accuracy of 73%.

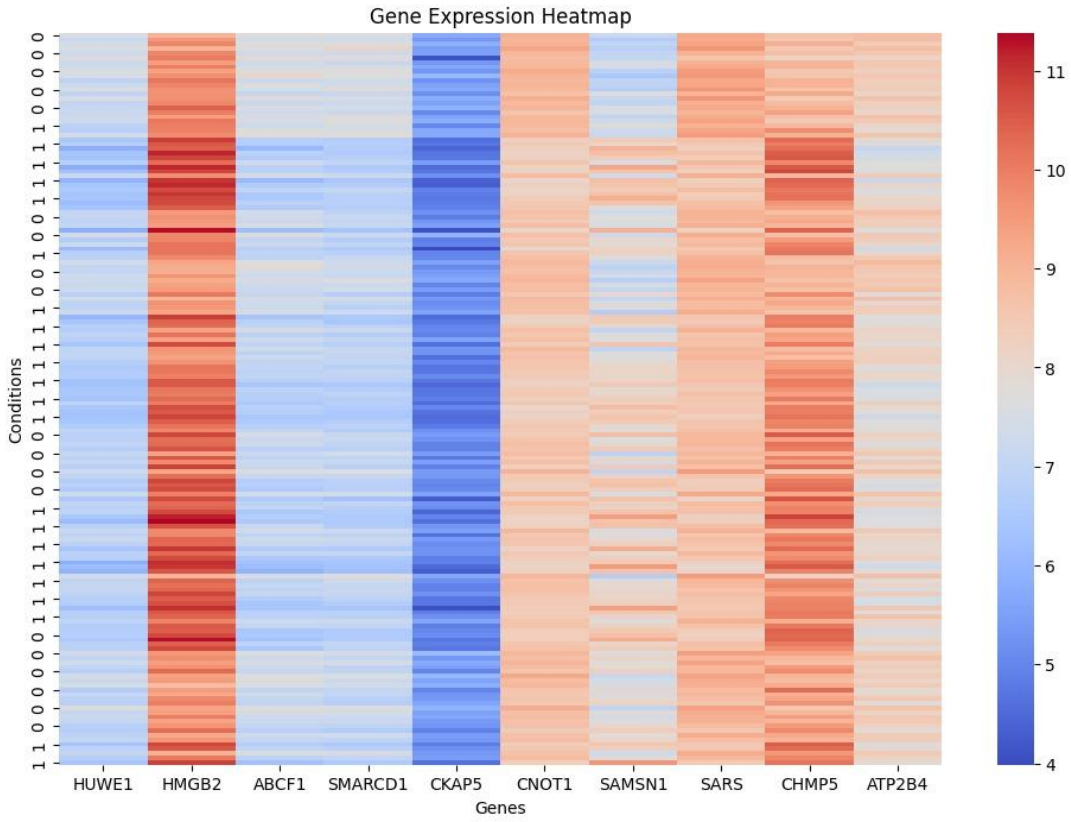
```
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy:.2f}")
```

```
Accuracy: 0.73
```

```
#Make predictions on the test data
y_pred = classifier.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy:.2f}')

# Get a classification report for more detailed metrics
classification_rep = classification_report(y_test, y_pred)
print(classification_rep)
```

```
Accuracy: 0.79
              precision    recall  f1-score   support

           0       0.80      0.84      0.82        19
           1       0.77      0.71      0.74        14

    accuracy                           0.79        33
   macro avg       0.78      0.78      0.78        33
weighted avg       0.79      0.79      0.79        33
```

Autoimmune diseases can lead to increases or decreases in gene expression levels, and different conditions may act differently.



We found some significant genes in the autoimmune diseases we studied, which included Rheumatoid Arthritis, Sjogren's Syndrome, Systemic Sclerosis, Systemic Lupus Erythematosus, etc.

- STAT1: The signal transducer and activator of transcription 1 (STAT1) is a critical player in the JAK-STAT signalling pathway. Abnormal activation of this pathway has been observed in SSc and is associated with fibrosis and immune dysregulation.
- IFI44: IFI44 is an interferon-inducible gene. In SSc, interferon pathways have been implicated, and increased interferon signalling is observed in some patients.
- LA-DRB1 (Human Leukocyte Antigen - DR Beta 1): This gene is a significant genetic risk factor for rheumatoid arthritis. Specific alleles within the HLA-DRB1 gene, known as the shared epitope, are strongly associated with an increased risk of developing RA.
- SMARD1: SMARD1 is a protein involved in chromatin remodelling, which regulates gene accessibility. Variations in SMARD1 have been associated with an increased risk of SS, particularly in individuals with specific HLA gene variants.
- SAMD9: Sterile Alpha Motif Domain-Containing Protein 9 (SAMD9) is a gene implicated in various developmental abnormalities. It is involved in growth regulation and cellular development. In the context of rheumatic diseases, SAMD9 may play a role in cellular processes relevant to immune dysregulation, as observed in conditions such as systemic sclerosis (SSc).
- UBE2V1: UBE2V1, part of the ubiquitin-proteasome system, regulates protein degradation. In rheumatoid arthritis (RA), UBE2V1 may influence inflammatory and immune responses by modulating protein turnover through the ubiquitin-proteasome system.

Specific genes implicated in these disorders include HLA-DRB1 for RA, specific polymorphisms within the HLA region for SjS, multiple genetic loci including HLA genes for SLE, and HLA and non-HLA genes like STAT4 and IRF5 for SSc. Additionally, variations in cytokine genes, such as TNF and IL-6, may contribute to the development of autoimmune responses. The interplay between these genetic factors and environmental triggers plays a crucial role in the onset and progression of autoimmune diseases, contributing to the dysregulation of the immune system. Environmental factors such as infections, hormonal changes, and exposure to certain substances can further stimulate inflammation and disrupt immune balance. Inflammation plays a central role in autoimmune diseases, causing damage to tissues and organs.

References

1) Mossotto, E., Ashton, J.J., Coelho, T. *et al.* Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. *Sci Rep* 7, 2427 (2017). https://doi.org/10.1038/s41598-017-02606-2
2) Srivastava A, Sathiyasekharan M, Jagadisan B, Bolia R, Peethambaran M, Mammayil G, Acharya B, Malik R, Sankaranarayanan S, Biradar V, Malhotra S, Philip M, Poddar U, Yachha SK. Paediatric inflammatory bowel disease in India: a prospective multicentre study. Eur J Gastroenterol Hepatol. 2020 Oct;32(10):1305-1311. doi: 10.1097/MEG.0000000000001859. PMID: 32796356.
3) Loddo, I. and Romano, C. (2015) 'Inflammatory bowel disease: Genetics, epigenetics, and pathogenesis', *Frontiers in Immunology*, 6. doi:10.3389/fimmu.2015.00551.
4) Database: https://adex.genyo.es/