A Stepwise Diagnostic Pipeline for IBD Classification Using Random Forests

Gavin Bulthuis

December 14, 2025

Abstract

Inflammatory Bowel Disease (IBD) diagnosis typically relies on a combination of clinical history, symptom assessment, laboratory biomarkers, and endoscopic evaluation. However, these data sources vary widely in availability, cost, and diagnostic utility. To address this, I developed a tiered multi-stage machine learning framework that progressively integrates groupings of features to predict the status of IBD in a given patient. Using a Random Forest classifier and stratified cross-validation, I evaluated model performance at each tier to quantify the incremental diagnostic value of more advanced features. This tiered approach provides a scalable and clinically aligned diagnostic pipeline that mirrors a real-world demonstration of patient care and cost decision making.

1 Introduction

1.1 Clinical Background

IBD continues to be a growing global health burden characterized by chronic inflammation in the gastrointestinal tract. The two main types of IBD can be classified as Crohn's Disease (CD) and Ulcerative Colitis (UC). The current gold standard for diagnosis is utilizing a colonoscopy with biopsies. While this often gives physicians the answers they need to make a proper diagnosis, it can be an expensive procedure. On top of this, it requires significant patient preparation and small risk.

In clinical practice, the process for diagnosis can be extremely inefficient. Patients presenting with non-specific gastrointestinal symptoms such as pain, fatigue, or an increase of stools often undergo a litany of tests before getting a chance to receive a referral for endoscopy. This can cause delay for patients who need treatment as fast as possible and burden healthcare systems with unnecessary procedures in healthy individuals suffering from other disorders.

1.2 Problem Statement

The core inefficiency in IBD diagnostics is the lack of a unified stepwise system that can aid in determining not only the diagnosis, but the severity before having to undergo invasive procedures. Clinicians must weigh the cost of missing a

diagnosis (false negative) against the cost of unnecessary invasive testing (false positives). Currently, there is no widely adopted algorithmic approach to this process that integrates a number of factors into making a data-backed diagnosis pre-endoscopy.

1.3 Research Questions

This study aims to develop and validate a stepwise machine learning pipeline for IBD classification. Specifically, seeking to answer the following questions:

1. Can I reasonably give a physician a machine learning model that can help aid in their decision making throughout the entire diagnostic process or pieces of it?
2. Can a stepwise machine learning model reduce the number of unnecessary and expensive tests without making unforgivable mistakes?
3. What features or groups of features contribute most towards patients having IBD or being healthy?

2 Data

2.1 Data Source

I utilized the public-facing IBD Multi'omics Database. There were 131 patients (104 IBD, 27 Non-IBD) that were part of the multi-omics study. The dataset includes clinical metadata, patient surveys, and biomarker/procedural findings.

Crohn's (CD)	Healthy (nonIBD)	Ulcerative Colitis (UC)
66	27	38

Table 1: Class Distribution (Condensed to IBD vs nonIBD for modeling)

2.2 Dataset Engineering

To mirror the real-world clinical workflow, features were grouped into four tiers of increasing complexity and cost. **Figure 1** gives examples of features you may see in the following tiers.

- **Tier 1: Risk Factors (Demographics & History)**
  The Baseline Susceptibility. This tier consists of static variables that describe a patient's background risk before any clinical assessment is performed. These are low-cost data points often collected during intake.
- **Tier 2: Patient-Reported Symptoms**
  The Clinical Presentation. This tier captures the patient's current symptoms. Unlike Tier 1, these features represent the active signs of disease but remain "cheap" and only require an office visit.

- **Tier 3: Biomarkers**
  The Laboratory Evidence. This tier introduces objective measurements of inflammation obtained through blood and stool samples. These are more expensive than symptoms but avoid the procedures.
- **Tier 4: Endoscopy & Histology**
  The Definitive Diagnosis. The final tier includes data that can only be obtained through procedures (colonoscopy/biopsy). This represents the "ground truth" but comes with the highest patient burden and cost.

Feature Engineering: A Stepwise Clinical Progression

Tier	Type	Features
Tier 1: Risk Factors	*Demographics & Patient History*	Weight, Height, BMI, Nationality, Education, Occupation, Premature Birth, C-Section Birth, Hospitalizations, Smoking Status, Anti-Inflammatory Meds, Contraception, Antibiotics, Alcohol Use, Red Meat, Fruit & Veg Consumption
Tier 2: Patient-Reported Symptoms	*Clinical & Current Health*	Recent Diarrhea, Well-Being, Depression, Bloody Stools, Stool Urgency, Weight Loss, Abdominal Pain, Nausea, Sleep Difficulty, Fever, Fatigue, Leisure Activity, Social Life, Eye Condition, Vomiting, Back Pain, Mouth Sores
Tier 3: Biomarkers	*Labratory Results*	C-Reactive Protein (CRP), ESR, Fecal Calprotectin, SCCAI, SIBDQ Score
Tier 4: Endoscopy & Histology	*Procedural Results*	Dysbiosis Score, Macroscopic Inflammation, Modified Baron's Score, SES-CD Score, Area Specific Severity, Ulcerations, Biopsy Location

Figure 1: Visualizing tier progression with example features

2.3 Data Pre-Processing

The raw dataset contained multiple rows per subject across different testing types that the subjects underwent. To be able to classify a patient as one with IBD or not, I needed to aggregate this long data down to one row per patient. To accomplish this, I averaged continuous variables and took the mode of categorical variables. Missing values were common in this dataset and were left alone due to the small sample size of patients and inability to estimate healthcare data. Ordinal variables were properly ordered to preserve their true directions during modeling.

2.4 Ethical Considerations

2.4.1 Data Privacy and Consent

This study utilized data with a lack of identification from the IBD Multi'omics Database. While all personal health information was kept anonymous, I acknowledge the importance of keeping patients' data secure and confidential throughout the entire analysis. No attempts were made to identify participants.

2.4.2 Algorithmic Fairness and Bias

A significant ethical limitation of this study is the sample size. With only 131 subjects, there is the potential for demographic homogeneity within the data. Machine learning models that are trained and tested on non-representative groups risk algorithmic bias and could cause misleading results when utilized in

different ethnic, social, or socioeconomic groups. If this approach was actually implemented in a practical setting, the model would need to be trained on a diverse dataset that doesn't under-represent any human being.

2.4.3 Model Utilization

Relying on a model to explicitly decide patient outcomes is not reasonable and will likely never be used as the "gold standard" in medicine. The aim of this project is to build an IBD classification model that merely aids in decision making and helps understand a patient's probability of having IBD. This assures that truly sick patients are not being undiagnosed.

2.4.4 Subject Distribution

The subjects in this study have some association with IBD or even other gastrointestinal issues and were not chosen at random. This is clear considering the difference in class size and purpose of the study.

3 Methods

3.1 Modeling Strategy

I implemented a Random Forest Classifier (using the 'ranger' package in R) due to it's robustness against overfitting, ability to handle missing values, and working through non-linear relationships. Most importantly, I chose this method because it simulates how a doctor thinks when diagnosing a patient. They can scale levels of decisions and add up factors to make their ultimate diagnosis. The unique aspect of this project was the cumulative adding of tiers into the model. I trained separate models on each tier of features and reasonable groupings of tiers to quantify the value of information being added at each stage of the diagnosis process.

3.2 Class Imbalance

As mentioned earlier, the dataset classes were significantly imbalanced with 79% of the participants falling in the IBD category with either a diagnosis of CD or UC. To prevent model bias towards the positive class, I used random under-sampling within cross validation folds. Each fold randomly sampled 27 IBD patients to have an even representation of classes when training the model. After the model was trained, the validation was performed on the entire dataset to reflect more realistic outcomes.

3.3 Evaluation Metrics

The performance of the machine learning model was evaluated using Area Under the Receiver Operating Characteristic Curve (AUC) and Balanced Accuracy (mean of Sensitivity and Specificity). AUC tells how likely it would be that the model ranked a patient with IBD over a healthy one. The Balanced Accuracy gives an estimate of how many predictions were correct by the model. I utilized this over standard accuracy due to the class imbalance. Given the clinical

context, I also wanted to look at the Positive Predictive Value (PPV) and Negative Predictive Value (NPV). Maximizing the NPV is critical for any medical context so that false negatives aren't made by missing diagnoses. This is why I will eventually simulate patients through the tiered system so that I can make up for low NPV due to the class imbalance.

3.4 Patient Simulation

To translate model probabilities into real-world clinical decision making I developed a stepwise simulation. The patients begin in the first tier and proceed to the next tiers only if the probability of having or not having the disease lands within a certain threshold. The threshold becomes less strict as I feed the model more information.

3.4.1 Resource Constrained Simulation

I further evaluated the diagnostic pipeline under more realistic hospital constraints. In reality, there are limited numbers of invasive testing that hospitals can undergo in a given set of time. To model this, I ran a different simulation where I added a constraint that only 15% of patients could be scoped. Patients who I was unsure of were ranked by their predicted risk in the previous tier and the scopes were given to the higher risk candidates until the capacity was reached. The remaining were assigned a diagnosis based on previous probability. This allowed me to calculate how well the model performed when forced to make more difficult decisions.

4 Results

4.1 Model Performance by Tier

Initially, I evaluated the diagnostic performance of the Random Forest classifiers by creating four models with incrementing features (1-4). Each model would aggregate the previous tiers' features until I reached the final model. While this works, I also wanted to build out models for each individual tier so that a physician could receive assistance at any stage of the diagnosis process. There are also likely scenarios in which some of the data may not be accessible, increasing the need for these individual models. In addition, I also created other small groupings of tiers that made sense such as only the "cheap" features like demographics and symptoms or only the "test-related" features like biomarkers and endoscopic findings.

All ten models were built using the same framework with model specific tuned hyperparameters. The performance was evaluated on the independent test set in each validation fold (**Figure 2**).

Model Performance Metrics

Model	AUC	Bala d_Acc	PPV	NPV	Prevalence
t2t3t4	0.923	0.816	0.944	0.537	0.794
t2t3	0.918	0.822	0.938	0.600	0.794
t2	0.904	0.830	0.946	0.579	0.794
t3t4	0.890	0.769	0.929	0.457	0.794
t1t2t3t4	0.884	0.817	0.937	0.583	0.794
t1t2t3	0.877	0.830	0.946	0.579	0.794
t1t2	0.843	0.774	0.923	0.500	0.794
t4	0.817	0.715	0.928	0.355	0.794
t3	0.812	0.717	0.905	0.404	0.794
t1	0.701	0.641	0.877	0.310	0.794

Figure 2: Model Performance Metrics Across All Tiers

4.1.1 Takeaways from Results

- **Symptoms Drive Prediction:** Patient-reported symptoms (Tier 2) had the most predictive power on their own and also contributed to each of the other 4 of the top 5 models. Tier 2 alone can help drive high probability predictions while saving more invasive testing for less confident cases.
- **The Missing Data Trap:** Surprisingly, the more "expensive" features included in Biomarkers (Tier 3) and Endoscopy & Histology (Tier 4) were not all that powerful on their own. There was a lot of missing data in Tiers 3 and 4, which likely contributed to their poor performance in comparison to other tiers. Despite the lower AUC's alone, they still are powerful in combination with other tiers.
- **Tier Efficiency vs Complexity:** Stacking all of the features proved to be unnecessary in most cases. Combining all features can create noise and make the process less cost-efficient, hence why I tested multiple subsets. The best model doesn't need Tier 1 because there is not a strong signal towards either class and there is simply a lot of noise within features in that tier. This same idea applies to adding Tier 1 to Tier 2 and 3 combined.
- **Screening Potential:** Despite being the worst performing model, Risk Factors (Tier 1) still can have moderate predictive capability and suggests that machine learning can successfully flag at-risk patients before needing a doctor.
- **The Standard Method Works:** It isn't the best overall performing model, but combining all of the tiers (t1t2t3t4) still ranks inside the top

5 in AUC and Balanced Accuracy making it useful in making end-to-end diagnosis predictions.

- **Cost Efficiency and Accuracy:** The Tier 2 and 3 (t2t3) model is the highest performer in terms of saving money while still avoiding false negatives. If a majority of patients presenting with IBD-like symptoms will be seen by a doctor at the symptom level and receive basic lab tests, why would this not be our best option? This model can be utilized and only lose 0.005 AUC and improve accuracy while saving money and eliminating the need for a patient to undergo an invasive colonoscopy.

4.2 Feature Importance

To interpret how the features were contributing towards predictions, I computed SHAP (SHapley Additive exPlanations) values for all of the models and looked at basic Random Forests feature importances as well. **Figure 3** outlines which features had the highest scaled importance when making predictions across the entire suite of models.

Figure 3: Top 5 Most Important Features by Model

The most influential features included `recent_diarrhea`, `education`, `dysbiosis_score`, and `fecalcal`. Outside of `education`, all of these align with clinical expectations as strong predictors of IBD.

To analyze the SHAP values across all models, I looked at them for the all inclusive model (t1t2t3t4) in **Figure 4** .

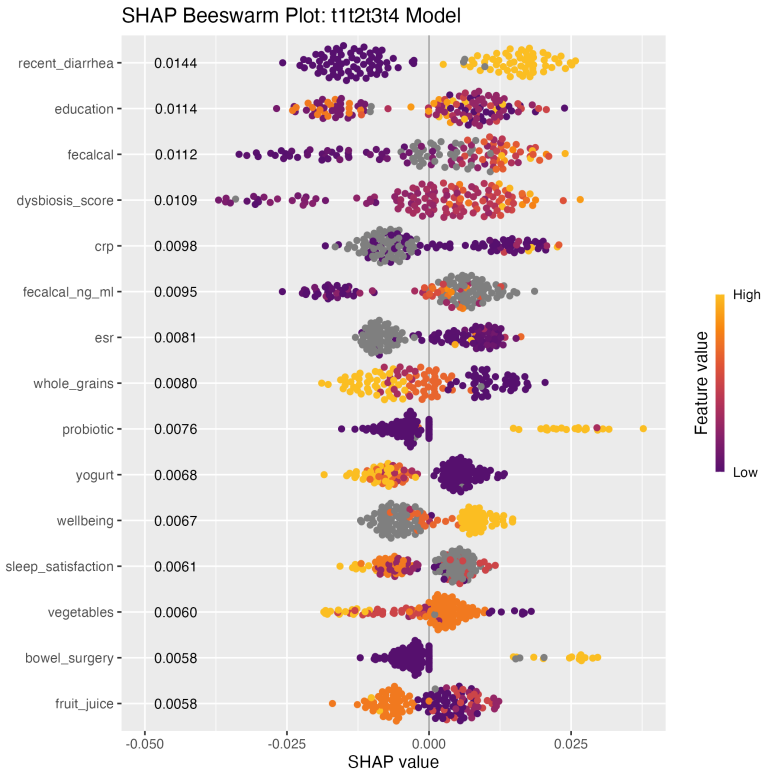

Figure 4: SHAP Plot for Model Using All Features

These tell more of a story on how these features directly contribute to the predictions. Most of them align with clinical expectations and show value regardless of what tier they were in.

- `recent_diarrhea`: Having more diarrhea coincides directly with having IBD or being healthy.
- `education`: There are many mixed signals when it comes to this feature, with no direct correlation with either class.
- `fecalcal`, `dysbiosis_score`: Low and high values directly contribute to being healthy or not.
- `probiotic`: People who take probiotics likely have had gastrointestinal issues before.
- `whole_grains`, `yogurt`, `vegetables`: Making healthier food choices can impact your health.
- `wellbeing`: Overall wellbeing can be a direct tell when it comes to being ill or diagnosing an illness.

4.3 Simulation Results

I implemented a simulation to model the clinical process of these classifiers. By applying strict probability thresholds and escalating uncertain cases to the next tier, the stepwise system was deemed functional.

4.3.1 Standard Incremental Tiers

The first iteration of the simulation involved starting at Tier 1 and adding the next set of features until I reached the Tier 4 (t1t2t3t4). This made the most sense clinically as the provider continues to receive data and adds it to previous information to make informed decisions.

The probability thresholds for this simulation were defined as:

- **Tier 1 (t1)**: $P(IBD) \geq 0.80$ = IBD or $P(IBD) \leq 0.20$ = nonIBD
- **Tier 2 (t1t2)**: $P(IBD) \geq 0.70$ = IBD or $P(IBD) \leq 0.30$ = nonIBD
- **Tier 3 (t1t2t3)**: $P(IBD) \geq 0.60$ = IBD or $P(IBD) \leq 0.40$ = nonIBD
- **Tier 4 (t1t2t3t4)**: $P(IBD) \geq 0.50$ = IBD or $P(IBD) < 0.50$ = nonIBD

After simulating, the overall process ended up making the correct prediction 91% of the time. In addition to this, the model was able to make the correct classification in 100% of cases before reaching Tier 4, with 60% of patients being handled before undergoing expensive testing through colonoscopy.

Table 2: Patient Handling and Accuracy by Diagnostic Tier

Stopped at Tier	Patients Handled	Tier Accuracy
Tier 1	4	1.000
Tier 2	37	1.000
Tier 3	38	1.000
Tier 4	52	0.769

The accuracy significantly dropped when handling the patients in Tier 4, but positively this is the point of the diagnosis process where a model would likely be disregarded as the invasive testing alone should give the clinician enough information to make a diagnosis.

4.3.2 Optimal Model Choice Simulation

Instead of only using the incremental feature tiers, I wanted to look at a simulation approach that still added features, but in a way that chose the best models.

Adding the Tier 1 features to any model tended to make it worse, so I will only use Tier 1 for the initial screening. For Tier 2, I will look at it alone as it was the best performing model that used only its own features. Adding Tier 1 features to Tier 2 features results in a decrease of 0.06 AUC. Next, the model combining Tier 2 and Tier 3 features had the second highest performance in terms of AUC. This will aid in making the tough decisions before sending a patient to the final diagnosis stage instead of using Tier 1, 2, and 3 which

performs significantly worse than simply using Tier 2 and 3 together. The final decision will come down to our best performing model (0.923 AUC) using Tier 2, 3, and 4 features together.

After running this version of the simulation with the same probability thresholds, I saw the accuracy jump up to 92%. The accuracy before reaching Tier 4 wasn't 100%, but still was 99.4% on average. I also saw improvements in the number of patients being handled before Tier 4, with an increase from 60% to 83%.

Table 3: Patient Handling and Accuracy by Diagnostic Tier

Stopped at Tier	**Patients Handled**	**Tier Accuracy**
Tier 1	4	1.000
Tier 2	58	0.983
Tier 3	43	1.000
Tier 4	26	0.654

The accuracy was worse in Tier 4 in this simulation, but I was able to make a significant increase in the amount of patients handled before this stage. I think the benefit of this outweighs the decrease in accuracy, especially when considering Tier 4 is the stage where the doctor makes a decision based on the endoscopic findings. This simulation provided better patient disease classification while also saving patients and hospitals money.

4.3.3 Sensitivity Simulation

Another simulation that I wanted to try was the scenario where I could only have a limited number of patients receive the invasive testing of Tier 4. I set the threshold that only 15% of patients (20 in our sample) were able to receive an endoscopic procedure. The simulation was done on the optimal models and same decision thresholds from the previous section.

If I ran this purely on the 131 patients, it would perform horribly due to the prevalence of the positive class (79%). I would send a lot of sick patients home. To make this more realistic, I used 200 patients with 40 of them actually having IBD. I created a synthetic sample from the original data with replacement to create more healthy patients. I allowed for 30 (15%) patients to get scopes if necessary.

This experiment performed extremely well with 96% of patients receiving the correct diagnosis overall. 81% of cases were handled before sending the top 30 cases to receive scopes. Of the 38 cases that weren't finalized prior to this tier, the top 30 (ranked by probability) received scopes and the remaining 8 were sent home despite having IBD. In reality, these 8 extras would likely be given treatment options or have their procedure deferred to a later date. I doubt that these patients would be disregarded and would still receive proper treatment at some point. Their symptoms were likely non-severe due to having a lower probability than the top 30 anyways.

Table 4: Distribution of Denied and Tier-Handled Cases

Category	Number of Cases
Denied Scope (Capacity)	8
Tier 1	18
Tier 2	111
Tier 3	33
Tier 4 (Scoped)	30
Total Missed Cases	**8**

The only patients who were predicted incorrectly were the ones who were denied scopes due to the capacity constraints. The numbers handled per tier need to be taken lightly as I was sampling a lot of Non-IBD patients from the original dataset. But, this does tell us that our model can be utilized at a larger scale when doctors are presented with actual IBD patients at a low rate when giving standard care. This answers the major question on whether this approach can safely rule out disease in lower prevalence settings.

5 Discussion

5.1 Clinical Implications

The primary contribution of this study is the validation of a stepwise machine learning diagnosis system for IBD. In current clinical practice, the decision to order an endoscopy is often subjective and requires an elongated process with increased costs for both hospitals and patients. This stepwise simulation demonstrates that machine learning can provide data-driven second opinions that can be used by doctors at any level of the diagnosis process.

By utilizing Tier 2 (Symptoms) and Tier 3 (Biomarkers) models, clinicians can confidently rule out disease in low-risk patients without resorting to invasive procedures. This optimized simulation showed that over 80% of patients could be diagnosed properly before undergoing endoscopy with a high accuracy. In a real-world health care system, this translates to significant cost saving and reduces wait time for patients who critically need endoscopic evaluation. Furthermore, the high NPV observed in our low-prevalence sensitivity simulation suggests that our modeling approach is safe for screening and minimizes the risk of dropping truly sick patients.

5.2 Choosing Tier Sets Wisely

A critical finding that I uncovered was the relationship between the volume of data (features) and model performance. Contrary to the assumption that having more data is better, it was observed that adding Tier 1 features (Risk Factors) to higher tiered models worsened performance. Specifically, the combined model with all of the features (t1t2t3t4) performed significantly worse than simply using the last 3 tiers (t2t3t4) which end up being the highest performing model.

This suggests that once active clinical symptoms have been found (Tier 2),

the static risk factors simply add noise rather than signal. This mirrors practice: if a patient presents with bloody stools and high fecal calprotectin, their education history or whether they were born via c-section become irrelevant.

I also observed that using the Tiers on their own still provided value, but overall were less valuable than grouping them. Outside of Tier 2, the worst three models by AUC were Tier 1, Tier 3, and Tier 4 isolated. Not having full context and information on the patient was clearly important when making predictions. It is vital to note that if there was more variance and overall more populated values for Tiers 3 and 4, there may have been increased performance as those are essentially directly correlated with the outcome.

5.3 Limitations

While promising, this study is subject to several limitations.

- **Sample Size**: The sample size is relatively small for machine learning applications, increasing the risk for overfitting despite the use of multiple cross-validation folds. Ideally, I would train this model on thousands of patients' data.
- **Class Imbalance**: With a 79% disease prevalence, this doesn't reflect the general population. I attempted to down sample the majority class, but this still could lead to misleading results. While the sensitivity simulation attempted to correct for this, I was still sampling from the same set of patients. External validation on a larger and more diverse sample would be necessary.
- **Missing Data**: The dataset contained a lot of missing values, particularly in the Biomarker and Endoscopic tiers. This could have hidden a lot of patterns that truly were predictive of IBD since these tiers are directly correlated with the positive class.
- **Data Imputation**: By averaging or taking the mode of values throughout a variety of visits, I could lose information that is vital in the patient's disease status. I was essentially estimating their one time visit data.
- **Study Bias**: With 104 out of 131 patients having IBD, these subjects were definitely not chosen at random and the "Healthy" class likely dealt with some version of a gastrointestinal illness. The study also is looking at patients who were already given a diagnosis. This model is trained as if I were seeing prospective patients presenting with IBD symptoms, so completely different patterns in random subjects could be seen. I would expect the pre-trained models to perform worse in practice because of this.

6 Conclusion

In conclusion, this study demonstrated that a stepwise Random Forest classifier can effectively quantify IBD risk without having to undergo invasive procedures. These findings indicate that symptoms (Tier 2) and basic biomarkers (Tier 3) hold the highest diagnostic value. Through an array of simulations, it was shown that the pipeline could theoretically decrease unnecessary colonoscopies

by prioritizing high-risk patients. The simulations achieved up to 96% accuracy in a low prevalence group even when resources were constrained. Doctors can use this array of models to aid in their decision making at any stage of the process. They can also view the feature analysis with SHAP to determine which values of features contribute most to either having IBD or being healthy. Ultimately, this framework offers a scalable and cost-effective blueprint for the diagnostic process that moves away from a "test all" approach.

7 Future Work

Future research should focus on key areas to translate these findings into practice:

- **Increase Data**: I would aim to gather the same information from thousands of patients and balance out classes for improved performance in a real clinical setting. This increase would also focus on random subjects that have no association to IBD. This way the model can be used to identify risk in unknown cohorts while our current set of models have some potentially biased data. It would be crucial to also ensure that this data is complete and doesn't have as much data missing.
- **External Validation**: I would need for our models to generalize well to completely unseen data. The cross-validation folds among 131 patients can only do so much, especially when having to completely down sample a class.
- **Long Term Analysis**: I could utilize the time-series nature of the raw data from the Multi'omics Database to analyze how the disease flares or progresses over time so that I can uncover patterns to potentially improve outcomes.
- **Model Integration**: I could build out a system that would allow a physician to input data for the features used and receive real-time probabilities and risk scores during patient screening.

Data and Code Availability

The raw data supporting the findings of this study are openly available in the IBD Multi'omics Database (https://ibdmdb.org/results). The data processing scripts, Random Forest modeling pipeline, and simulation code used for this analysis are available by request and may be publicly available on my GitHub (gavin-bulthuis).

Acknowledgements

I'd like to thank the IBDMDB investigators and the participating patients for making this multi-omics dataset publicly available. This research was conducted as part of the DSCI 4093 Capstone Project requirements at the University of Minnesota. Thank you to Professor Julian Wolfson for all of your support and

guidance during the supervision of this project; it could not have been a success without your help!

References

[1] Greenwell, B. (2021). *fastshap: Fast Approximate Shapley Values*. R package version 0.0.7. URL: https://CRAN.R-project.org/package=fastshap.

[2] Zhu, H. (2021). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.3.4. URL: https://CRAN.R-project.org/package=kableExtra.

[3] Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758), 655–662.

[4] Robin, X., Turck, N., Hainard, A., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 77.

[5] R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[6] Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17.

[7] Wickham, H., Averick, M., Bryan, J., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686.

Appendix

Hyperparmeter Tuning Grid:

The following grid search was performed to optimize the Random Forest Models at each tier.

- **Number of Trees**: 250, 500, 1000, 1500, 2500, 2500
- **Mtry**: 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
- **Minimum Nodes**: 2, 4, 6
- **Splitting Rule**: Gini Impurity

Full Feature List by Tier:

Tier 1 Features

- **Demographics and Clinical History:**
  - `diagnosis`, `patient_weight`, `height`, `bmi`

  - hispanic_latino, non_usa_country_origination
  - education, occupation
- **Surgical and Medical History:**
  - cholecystectomy, abdominal_surgery, appendectomy, tonsillectomy
  - recent_hospitalization, endoscopy
- **Early-Life and Childhood Factors:**
  - farm_child, daycare_child, premature_child, cigarette_child
  - hospital_born_child, c_section_child, usa_child
  - breastfed_child, antibiotics_child, hospitalization_child, pets_child
- **Lifestyle and Medication Use:**
  - smoker, marijuana, marijuana_frequency
  - acid_reducers, acid_reducers_frequency
  - anti_inflammatory_meds, anti_inflammatory_meds_frequency
  - bile_production
- **Reproductive Health:**
  - female_contraception, female_contraception_type, female_contraception_brand
  - current_antibiotics, chemo
- **Dietary Intake:**
  - meat_preferences, sugary_drinks, artificially_sweet_drinks, fruit_juice
  - water, alcohol, yogurt, dairy, probiotic
  - fruits, vegetables, beans, whole_grains, starch
  - eggs, processed_meat, red_meat, white_meat
  - shellfish, fish, tea_coffee, sweets

Tier 2 Features

- **Gastrointestinal Symptoms:**
  - abdominal_pain, gas, nausea, vomiting
  - day_stools, night_stools, bloody_stool, stool_urgency
  - recent_diarrhea, decreased_appetite, weight_loss
  - abcesses_abdomen, abcesses_buttocks, fistula, draining_fistual
- **Systemic and Extraintestinal Symptoms:**
  - fever, fatigue, night_sweats
  - eye_condition, mouth_sores, skin_problems
  - arthralgia (joint pain), back_pain
- **Sleep Quality and Disturbance:**

  - sleep_difficulty, sleep_quality, sleep_satisfaction
  - sleep_refreshing, sleep_effort, sleep_problems
  - restless_sleep, sleep_worried
- **Mental Health and Quality of Life:**
  - depression, recent_depression
  - wellbeing, recent_wellbeing
  - social_life, leisure
- **Other History:**
  - recent_antibiotics

Tier 3 Features

- **Inflammatory Biomarkers:**
  - crp (C-Reactive Protein), esr (Erythrocyte Sedimentation Rate)
  - fecalcal, fecalcal_ng_ml (Fecal Calprotectin)
- **Clinical Indices:**
  - sccai (Simple Clinical Colitis Activity Index)
  - sibdq_score (Short Inflammatory Bowel Disease Questionnaire)

Tier 4 Features

- **Endoscopic and Histologic Scores:**
  - dysbiosis_score, is_inflamed
  - modified_barons_score (UC Severity)
  - ses_cd_score (Crohn's Disease Severity)
- **Segmental Severity Grading:**
  - ileum_severity, rectum_severity
  - right_colon_severity, transverse_colon_severity, left_colon_severity
- **Ulceration Characteristics (Size and Extent):**
  - ileum_ulcers, rectum_ulcers
  - right_colon_ulcers, transverse_colon_ulcers, left_colon_ulcers
  - ileum_ulcerated, rectum_ulcerated
  - right_colon_ulcerated, transverse_colon_ulcerated, left_colon_ulcerated
- **Inflammation Extent:**
  - ileum_inflammation, rectum_inflammation
  - right_colon_inflammation, transverse_colon_inflammation, left_colon_inflammation
- **Procedural Metadata:**
  - biopsy_location
  - recent_colonoscopy, bowel_surgery

Other Supporting Visuals

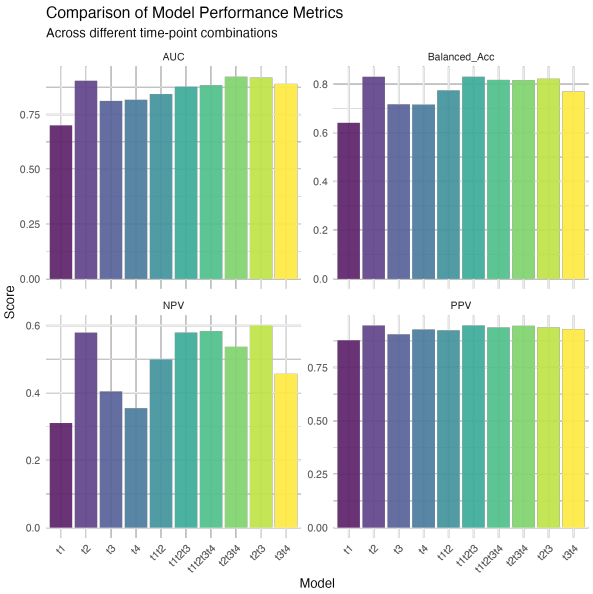

Figure 5: Model Performance Metrics

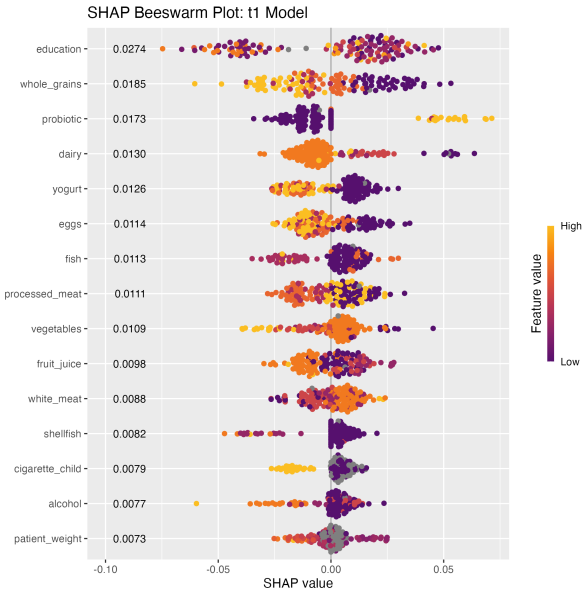

Figure 6: Tier 1 SHAP Values

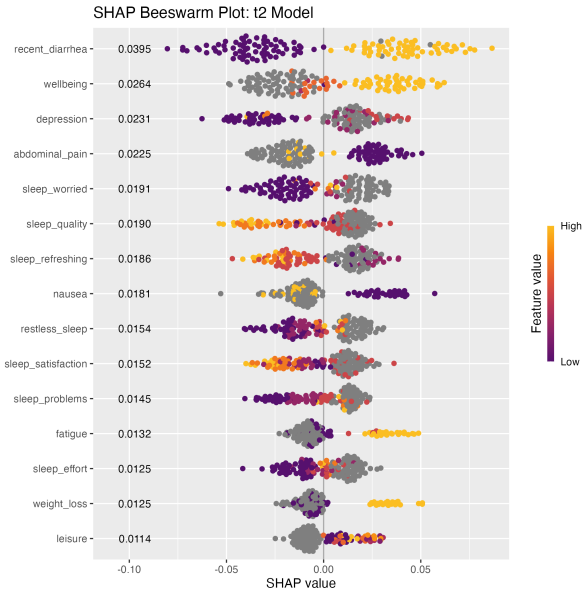

Figure 7: Tier 2 SHAP Values

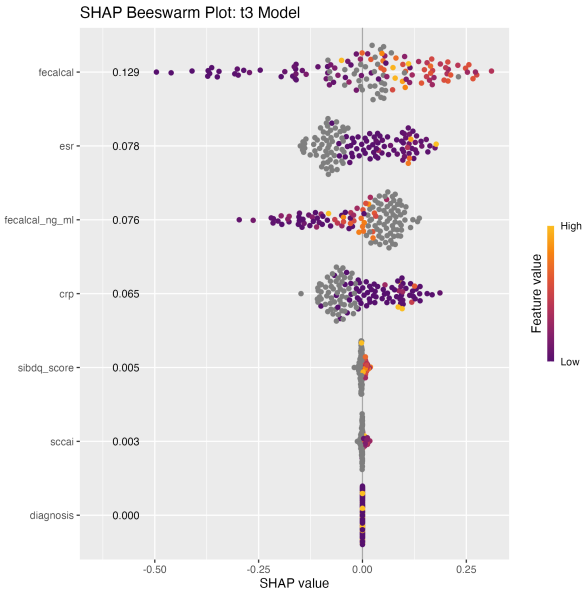

Figure 8: Tier 3 SHAP Values

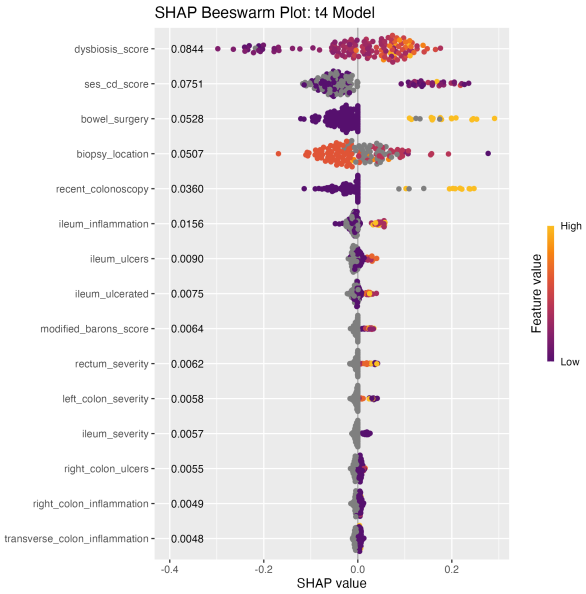

Figure 9: Tier 4 SHAP Values

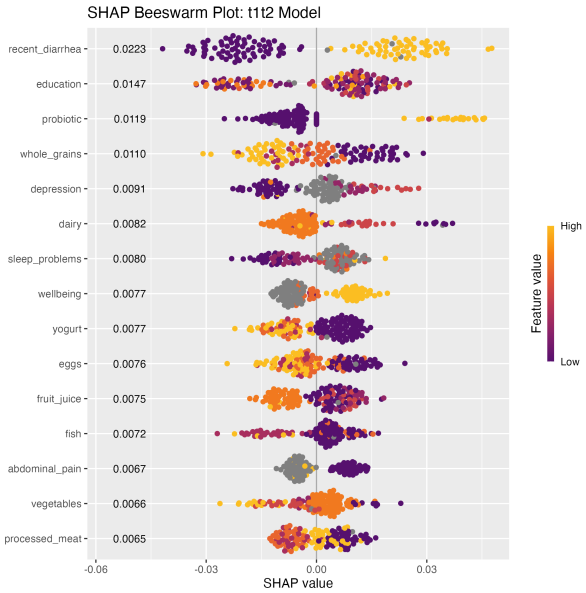

Figure 10: Tier 1 + 2 SHAP Values

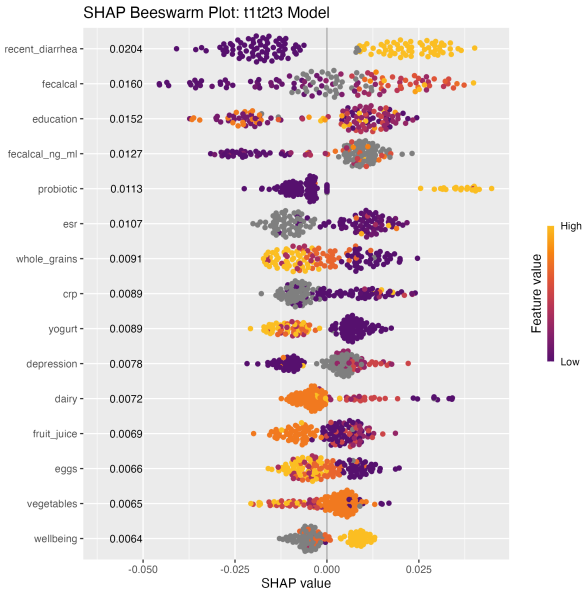

Figure 11: Tier 1 + 2 + 3 SHAP Values

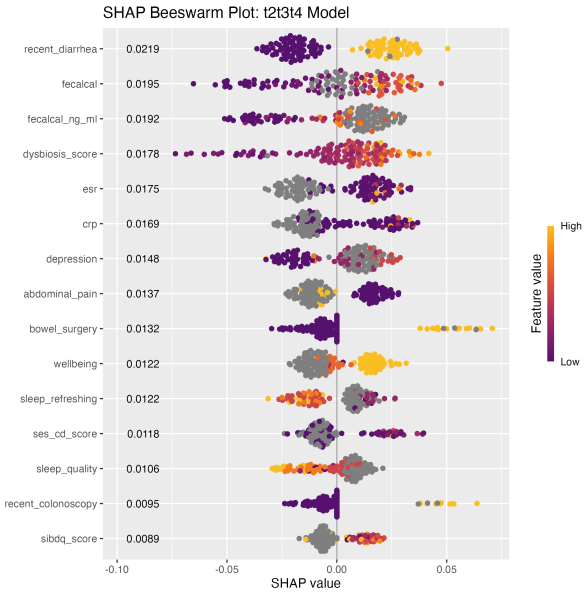

Figure 12: Tier 2 + 3 + 4 SHAP Values

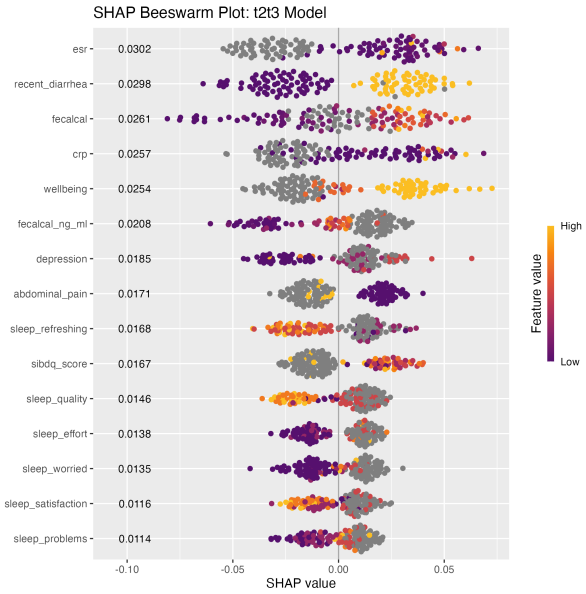

Figure 13: Tier 2 + 3 SHAP Values

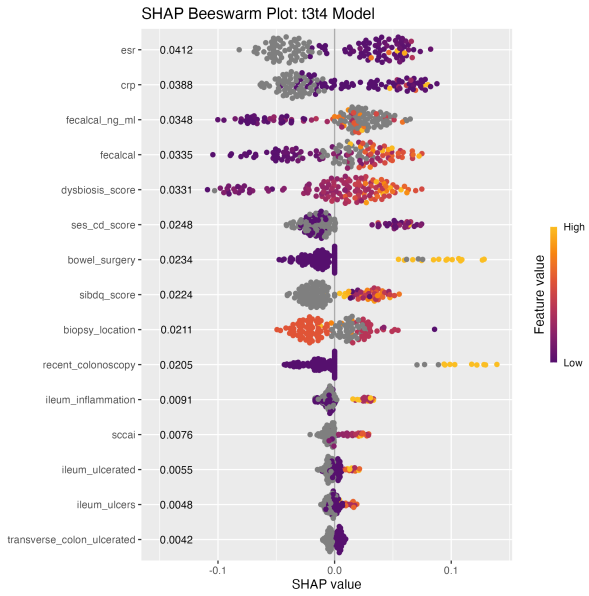

Figure 14: Tier 3 + 4 SHAP Values