IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Predicting the Prognosis of Ulcerative Colitis with Machine Learning

Aayushi Saxena

The Lawrenceville School, 2500 Main St., Lawrenceville, New Jersey, 08648, USA; aayushi.saxena.me@gmail.com
Mentor: Rosalyn Abbott

ABSTRACT: Ulcerative Colitis is a chronic inflammatory bowel disease with unpredictable progression, often requiring invasive and costly procedures (e.g., colonoscopy) for monitoring. This study developed a Support Vector Machine classification model that integrates histological information (i.e., clinically established biomarkers), treatment history, lifestyle factors, and demographic variables to classify UC severity. First, a prototype was developed using a synthetic dataset of 50 simulated patient profiles containing age, calprotectin levels, and disease severity scores. The pipeline successfully classified unseen inputs, confirming basic functionality. This prototype was migrated to a clinician-facing web platform, which allows users to upload datasets, determine a train-test split, select the target variable, and visualize performance metrics. Validation with a publicly available biomedical drug-classification dataset yielded 90.0% accuracy and 91.3% precision, demonstrating system reliability. These results show that the platform can manage end-to-end workflows in a scalable and interpretable way. Once UC-specific patient datasets are incorporated, the system will be expanded to enable clinicians to input live patient data and receive severity predictions and relapse risk estimates. By considering a variety of factors, this tool has the potential to reduce reliance on invasive procedures, improve monitoring efficiency, and advance precision medicine approaches for UC management.

KEYWORDS: Computational Biology and Bioinformatics, Computational Biomodelling, Machine Learning, Ulcerative Colitis, Prognosis.

Introduction

Ulcerative Colitis (UC) is a chronic, idiopathic inflammatory bowel disease (IBD) that causes continuous mucosal inflammation in the colon and rectum.[1] While some risk factors, such as diet, stress, genetics, and microbiota, are correlated with inflammation, UC remains largely unpredictable in its progression. Associated symptoms (including bloody diarrhea, tenesmus, and abdominal pain) typically overlap with other gastrointestinal (GI) disorders, necessitating diagnostic tools such as colonoscopy, biopsy, and imaging, which are often invasive, expensive, and not practical for frequent monitoring.[2]

Despite advancements in treatment, patients remain at high risk for relapse, complications, and long-term consequences such as toxic megacolon and colorectal cancer.[3] Non-invasive biomarkers are increasingly used to monitor inflammation, but as stand-alone tools, they lack the specificity and sensitivity required to reliably predict prognosis.

Recent literature has explored novel biomarkers and applied machine learning techniques to IBD, yet relatively few studies focus specifically on prognosis prediction in UC using clinically relevant biomarkers in combination with demographic and lifestyle data. Moreover, many existing studies emphasize model development without addressing how such tools can be translated into practical platforms with clinical relevance and usability.[4]

This research addresses these gaps by developing a Support Vector Machine (SVM) model to classify UC severity using a combination of biomarker, histological, demographic, and medication history data. A web-based platform designed for practical use by clinicians and researchers was developed. The platform enables users to upload patient datasets, select the target column/outcome (i.e., severity score), and define how the dataset is divided between model training and validation (e.g., an 80/20 split means 80% of the data is used to train the model, while the remaining 20% is used to test its accuracy). Results are then displayed immediately through interactive charts and performance metrics (such as accuracy, precision, F-1 score, recall, etc.), making the outputs both interpretable and clinically relevant. While UC-specific patient data is still being acquired, the system has already been tested with publicly available biomedical datasets to demonstrate end-to-end functionality and scalability.

Ultimately, this research developed a clinician-facing web platform that includes a validated machine learning model for UC prognosis and a framework for integrating multidimensional patient data into precision medicine tools. This project offers a pathway from computational modeling to clinical applications in UC care, such as improved prognosis.

Methods

This project followed a multi-step research and development process to design and implement a machine learning model for predicting the prognosis of UC.[5]

Step #1: Literature Search and Variable Identification:

In order to select an appropriate machine learning model, it was first necessary to identify and define the types of variables that best represent the clinical and biological complexity

DOI: 10.36838/IJHSR87.65

of UC, which arises from a complex interplay between the gut microbiota, the mucosal immune system, and host genetics.[6,7] To capture this biological complexity, variables were selected to reflect both molecular indicators of inflammation and individual-level differences that influence immune function and disease severity.

The variable selection process was refined in consultation with IBD specialists (Dr. Soula Koniaris and Dr. Sarag Boukhar), who emphasized the central role of biomarkers in diagnosing and monitoring UC. Based on their guidance, four clinically significant serum and fecal biomarkers (C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), calprotectin, and lactoferrin) were prioritized during the extensive literature search process. These biomarkers are routinely used to assess inflammatory burden and disease activity, providing quantifiable inputs well-suited for machine learning models. Additional demographic and lifestyle variables (including age, sex, diet, smoking status, and hospitalization history) were included to account for patient-specific variation in immune response and treatment outcomes.

Integrating patient-level characteristics with biomarker data produced a multidimensional feature set capturing the biological, behavioral, and demographic factors that shape UC progression. The final dataset combines biomarker values, clinical severity indices (e.g., Mayo Score, UCEIS), electronic medical records (diagnoses, history, problems, medication usage, lab test results), clinical data (relevant lab scores for biomarkers, medication journey, extraintestinal manifestations, disease phenotypes and locations, disease severity scores, etc.), and other demographic variables.[9] This comprehensive structure captures both systemic inflammation and individual variability, providing a strong foundation for developing and validating the machine learning model.

Step #2: Model Selection and Algorithm Capacity:

Because UC severity is influenced by a combination of immune, microbial, genetic, and lifestyle factors, a non-linear modeling approach was required to capture the complexity of these interactions. Based on the variable types identified earlier, both regression and multiclass classification frameworks were initially considered. Regression models predict continuous numeric outcomes (e.g., severity scores along a spectrum), whereas classification models assign cases to discrete categories (e.g., mild, moderate, severe).[10] Given that UC severity is typically evaluated through clinical categories, a classification approach was determined to be more appropriate.

Three supervised learning algorithms were evaluated: logistic regression, decision tree, and Support Vector Machine (SVM).[11] Logistic regression provides interpretability but is limited in modeling nonlinear biological relationships. Decision trees can model non-linearity but often overfit smaller or unbalanced datasets, reducing generalizability. The SVM algorithm was ultimately selected for its robustness in handling high-dimensional, non-linear, and limited-sample data, making it well-suited for biomedical contexts.[12]

To test the feasibility of this SVM model, an initial prototype was built in Python (Jupyter notebook), using the SVM model from the scikit-learn (sklearn) library, and trained on a synthetic dataset generated in Excel. The dataset consisted of 50 simulated UC patient profiles, generated based on clinically recognized biomarker ranges and severity indices reported in UC literature.[8] Each profile included two predictor variables—age and calprotectin level (μg/g)—and one target variable, disease severity, labeled on a 0–3 scale (0 = normal to 3 = severe). The SVM algorithm from the sklearn library was trained to map the predictor variables (defining the x vector as [age, calprotectin level]) to the output or target column (y = severity score). Although this dataset was artificial and limited in scope, it was instrumental in understanding how the dataset should be formatted later and allowed us to verify that the pipeline could successfully classify new inputs. Thereby, this step established a functional and replicable framework for future training on validated UC datasets.

Step #3: Prototype Development and User Testing:

To demonstrate user interaction and early data exploration workflows, an interactive prototype was developed using an R Shiny App (source code: https://github.com/aayushi-saxena-me/shinyapp). The app (hosted at https://aayushisaxena.shinyapps.io/shinyapp) allows for dataset/CSV upload, exploratory data analysis (summary statistics tables (with minimum, Q1, median, mean, Q3, and maximum values) and a one-sample hypothesis test), and a k-means clustering plot (a type of unsupervised machine learning algorithm). While this prototype provided proof-of-concept for user engagement, it lacked SVM integration and was not optimized for high-volume or clinical-grade deployment due to scalability limitations with the Shiny App framework.

Step #4: Platform Migration & Full Stack Development:

To overcome these limitations, the system was migrated to a full-stack architecture using Python/Django (source code: https://github.com/aayushi-saxena-me/statistical-analysis-app), with code development supported by Cursor (AI-powered IDE) and deployment via AWS CodeDeploy. The resulting web platform (https://statisticalanalysis.org) allows the user to upload patient datasets, define train-test splits, select the target/output variable, enable SVM learning analysis, and view performance metrics of the SVM classification, among other things. This architecture ensures scalability, reproducibility, and accessibility, making it possible for both researchers and clinicians to use advanced machine learning tools without requiring extensive programming expertise.

Step #5: Validation with External Datasets:

Since UC-specific datasets are still pending, platform validation was performed using a publicly available biomedical dataset (https://www.kaggle.com/datasets/prathamtripathi/drug-classification). This dataset was uploaded in CSV format, drug type was selected as the target column, RBF was selected as the SVM kernel, and a 20% Test / 80% Train split was selected. When the machine learning analysis was run, the SVM model achieved 90.0% classification accuracy, demonstrating the platform's ability to handle the full workflow (from data

upload to training and testing, to performance reporting) in a reliable and interpretable way. This confirms that the system works as intended and is ready to be applied to disease-specific data.

Once UC datasets are acquired, the same process can be repeated to produce clinically reliable predictions. Once trained on this UC dataset, in further iterations of the model, clinicians will be able to input new patient data and receive real-time outputs. For now, the focus is on demonstrating core functionality and evaluation workflows, not providing live patient predictions. Evaluating the model on new, unseen data provides a realistic measure of how well it will perform in real-world settings, an essential step toward developing a clinically dependable tool.

Results and Discussion

Literature Search: Biomarkers:

One of the initial focuses of this project was identifying biomarkers that could help predict the prognosis of Ulcerative Colitis (UC). While biopsy and endoscopy are effective diagnostic methods, their invasive nature and high cost limit their use in routine monitoring. Although some non-invasive imaging techniques have shown high diagnostic accuracy, they also require experienced personnel and sophisticated equipment, which can be costly and difficult to access regularly.[13]

Biomarkers, on the other hand, are medical signs different from symptoms that essentially act as measurable bodily substances which can provide information about both physiological and pathological conditions. These biomarkers can be found in various biological samples (e.g., blood, urine, tissues, etc.).[14] Some basic examples include vital signs from pulse and blood pressure to other serum, genetic, or inflammatory biomarkers, which require more complex laboratory tests. While biomarkers vary in specificity and sensitivity, they represent a practical alternative for monitoring disease progression. Through an extensive literature review and consultation with gastroenterologist Dr. Soula Koniaris, four clinically established biomarkers were prioritized: C-reactive protein (CRP), fecal calprotectin, lactoferrin, and erythrocyte sedimentation rate (ESR).

CRP is a well-known serum biomarker produced in the liver and elevated during acute-phase inflammatory responses. It has a relatively short half-life, making it a more dynamic indicator of inflammation. Although more sensitive for Crohn's disease than UC, CRP remains a useful indicator in UC.[15] Calprotectin, a calcium- and zinc-binding protein, is highly prevalent in phagocytic leukocytes (especially neutrophils) and constitutes about 60% of their cytosolic protein content. Neutrophil infiltration of the GI tract is a hallmark of UC, making calprotectin a strong indicator of active disease.[16] Fecal biomarkers such as calprotectin are particularly important in IBD diagnosis and prognosis because they reflect mucosal inflammation more directly. Fecal lactoferrin is another neutrophil-derived protein present in stool, and is significantly elevated in pediatric patients with active intestinal inflammation and IBD (for both UC and Crohn's disease).[17] ESR is a blood test that can indicate if there is inflammation (the immune system's response to injury, infection, and conditions such as immune system disorders (i.e., IBDs)) in your body. Erythrocytes are red blood cells, and the sedimentation rate is the time it takes for red blood cells to settle at the bottom of a test tube.[18] Though an elevated ESR is a less specific indicator of inflammation, compared to fecal lactoferrin, the biomarker still offers additional context as markers of systemic inflammation and neutrophil activity. According to Dr. Koniaris, elevated levels of these biomarkers described previously are commonly used in practice to assess whether inflammation is due to an intercurrent illness or not, and how to adjust their medication accordingly to prevent escalation of disease severity.

In addition to clinically validated biomarkers, emerging research highlights the potential of novel, more specific indicators. Several recent studies have used bioinformatic and machine learning approaches to identify differentially expressed mRNAs (DEMs) associated with immune activity in UC. One such study identified 216 DEMs related to neutrophil infiltration and leukocyte migration, with MST1L, OLFM4, and DPP10 standing out as potential prognostic markers.[19] Although these are not yet validated for clinical use, including them in future iterations of the model may help test their diagnostic utility.

Incorporating a wide range of biomarkers (established and experimental) along with basic patient information (e.g., age, gender, hospitalization history, diet, smoking, etc.) allows the model to account for the individualized nature of UC. Factoring these nuances increases the likelihood of building a tool that reflects real-world complexity and variation in disease behavior.

At this stage, we have a curated set of biomarkers and patient variables (including demographics) that will serve as the core features for the machine learning model. These inputs are both clinically meaningful and widely available, ensuring that the eventual tool will be feasible for real-world use while leaving space for incorporation of experimental biomarkers as the field evolves.

Computational & Model Selection:

Due to the complexity and non-linearity of UC progression, a Support Vector Machine (SVM) was selected as the primary model for classification. SVMs are a set of supervised learning algorithms effective for classification, regression, and outlier detection. They work by identifying a hyperplane that best separates data into classes, using kernel functions to project input into higher-dimensional space when necessary for non-linear separation.[20,21]

In this project, the model's input features include various clinical and biomarker data points, while the output is a severity score derived from UC scoring systems such as the Ulcerative Colitis Endoscopic Index of Severity (UCEIS) and the Mayo Score. For example, UCEIS scores range from 0-8, with higher scores indicating greater disease severity. These scores serve as the "y" values (labels) for training, while the "x" values are multidimensional input vectors consisting of variables like CRP, calprotectin, ESR, and patient demographics.

DOI: 10.36838/IJHSR87.65

SVMs require labeled training data, making them ideal for our objective: to classify the course and severity of UC into defined categories (e.g., mild, moderate, severe). Once trained, the model could help streamline clinical decision-making by suggesting potential treatment pathways or adjustments based on predicted risk of relapse.

Test Run and Prototype Tool:

To evaluate the viability of a machine learning model for UC prognosis, a Support Vector Machine (SVM) algorithm from the scikit-learn library was tested in Python.[22] The goal was to establish a working proof of concept while awaiting real clinical data. A synthetic dataset was created in Excel containing 50 hypothetical patient profiles. Each profile included two features—age (in years) and levels of fecal calprotectin (μg/g)—and was labeled with a correlating severity score from 0 to 3, representing disease intensity (mild to severe). This dataset was used to train the SVM in a Jupyter Notebook environment, and then the model was asked to classify new, unseen inputs using this training data.

For example, a 16-year-old patient with a calprotectin level of 400 μg/g was labeled as a 2 (moderate severity) in the dataset. When we asked the model to predict this same scenario ("clf.predict([[16, 400]])"), a score of 2 was returned as modeled in Figure 1. This prediction demonstrated that the model was functioning as intended, at least within the bounds of the simulated dataset.

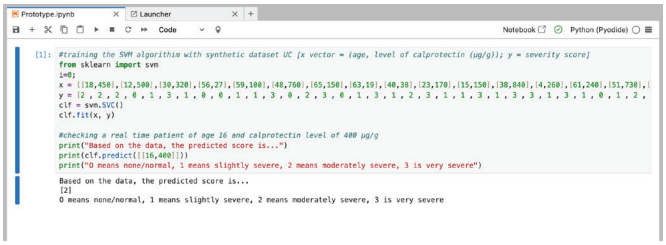

Figure 1: Implementation of a Support Vector Machine (SVM) in a Jupyter Notebook environment using a synthetic dataset of ulcerative colitis patient profiles. The model, trained on age and calprotectin level as predictive features, classified the test input [16, 400] with a severity score of 2, corresponding to moderate disease activity.

Although the dataset was small and artificially constructed, it validated the SVM's basic mechanics and confirmed the model pipeline, providing a foundation for interface and backend development. Future performance is expected to improve significantly with access to larger, clinically validated datasets.

Platform Development and External Dataset Validation:

In parallel to the Python-based modeling, an initial prototype was built using R Shiny and deployed at https://aayushisaxena.shinyapps.io/shinyapp to support clinical and user-facing interaction. This app enables users to upload CSV files, run summary statistics, perform hypothesis testing, and conduct unsupervised machine learning via k-means clustering. Though initially populated with a default brain tumor dataset, it can be used with any biomedical dataset. While the Shiny App was instrumental in shaping the user interface and front-end interaction, it could not run SVM models and would not scale effectively in a clinical setting.

To address the limitations of the prototype, a web-based platform (available at statisticalanalysis.org) was developed using Python/Django, migrated by Cursor, and deployed through AWS CodeDeploy to provide a unified environment for exploratory data analysis, statistical computation, and machine learning. Users can upload CSV or Excel datasets or generate sample data directly within the interface. Upon uploading, users can preview the data input and examine the key dataset information (number of rows, columns, and variable types) automatically produced by the system. Built-in visualization tools allow users to generate distribution plots, box plots, Q–Q plots, and correlation matrices, supporting a clear understanding of variable relationships and data distributions. The user can adjust which variables are the target columns in the analysis configuration on the left bar. The platform also computes descriptive statistics such as mean, standard deviation, variance, and percentile values, along with results from hypothesis testing modules, including t-tests and normality checks.

The platform also features a dedicated Machine Learning module that integrates an SVM classifier. Users can define the target column, choose a kernel function (e.g., Radial Basis, sigmoid, etc.), and specify the train-test split ratio (e.g., 80/20). The same number and type of input variables were used in both the training and validation datasets to maintain structural consistency and ensure the reliability of the model's evaluation. Once trained, the system automatically reports performance metrics (accuracy, precision, recall, and F1-score) along with a confusion matrix and performance summary chart for interpretability. While the prototype does not automatically normalize or standardize input data, users can manually format or preprocess their datasets prior to upload. This ensures flexibility for users who wish to control their own data preparation. Future iterations can be expanded to include built-in feature scaling to further optimize SVM performance and simplify the user workflow. The interactive interface allows users to refresh or rerun analyses without coding, making the platform suitable for both researchers and clinicians seeking to evaluate dataset trends or model outcomes efficiently.

To test this system, a publicly available biomedical dataset on drug classification was used. The dataset contained 200 samples (rows) and five drug classes as the target variable, with 5 input features: age, sex, blood pressure, cholesterol, and Na-to-K ratio.[23] The top of Figure 2 shows the performance metrics of the SVM classifier once it was trained (on 80%) and validated (with the remaining 20%) on the drug classification dataset. It produced an accuracy of 90.0% and precision of 91.3%, indicating that the SVM achieved balanced and reliable classification performance across all drug classes. The bottom half of Figure 2 shows another visualization of the performance metrics as well as the accompanying Confusion Matrix.

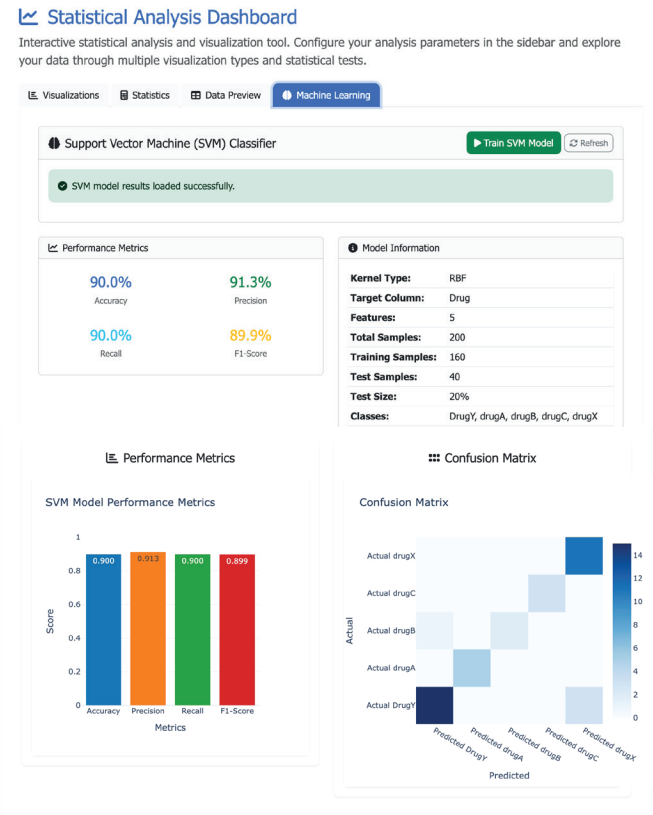

Figure 2: Web-based Statistical Analysis Dashboard demonstrating Support Vector Machine (SVM) model results on a drug classification dataset. At the top of the image, the dashboard interface displays model configuration and performance summary. On the bottom, the model outputs associated performance metrics and a confusion matrix showing classification accuracy (90.0%), precision (91.3%), recall (90.0%), and F1-score (89.9%).

While these datasets are not UC-specific, they effectively demonstrate the model's structure and performance. The architecture is ready to accept UC patient datasets, and once trained and validated on UC-specific data, the platform will not only evaluate predictive accuracy but also highlight which biomarkers and clinical variables most strongly influence disease prognosis.

This will enable refinement of the model by distinguishing between critical predictors and potential sources of noise. After validation, the system can expand beyond classification into a real-time decision-support tool, where clinicians may input patient-level data and immediately receive severity scores alongside probabilities of relapse, hospitalization, or treatment escalation. Producing performance metrics and actionable risk predictions for clinicians ensures that the platform is both scientifically robust and translationally relevant.

Limitations:

This study represents both a proof of concept and an applied framework for bridging innovation in machine-learning-based UC research with real-world clinical use; however, several limitations qualify its findings. First, the current system was developed on a synthetic dataset and validated on a drug-classification dataset that differs from UC in features and outcomes, so reported performance may not reflect UC prediction in practice. Second, the web app supports offline evaluation only: users upload a dataset, the platform performs a fixed 80/20 train–test split, and it reports metrics (accuracy, precision, recall, and F1 score) with a confusion matrix; per-patient, real-time prediction is not yet implemented. Preprocessing steps (e.g., feature scaling/normalization) are not fully specified, which matters for SVM stability and reproducibility. Before clinical use, the platform also requires testing and documentation for data security, scalability, and usability on UC-specific, multi-institutional datasets.

■ Conclusion

This study demonstrates the feasibility of applying machine learning to predict the prognosis of UC by integrating clinically established biomarkers, demographic variables, and treatment history into an SVM model. Through an extensive literature review and consultation with IBD experts, a relevant set of features was identified and used to build a classification system that distinguished UC severity in simulated data. This system was then migrated to a web-based platform and validated using an external biomedical dataset, showing reliable performance across key metrics including accuracy, precision, recall, and F1 score.

Overall, these results establish a functional proof of concept that bridges computational modeling with practical application. Moving forward, future implementations will involve incorporating UC-specific clinical datasets directly into the platform to train and evaluate the model on real patient data. This process will include structured data acquisition through clinical collaborations, evaluation of noisy or missing variables, and the development of automated preprocessing pipelines for normalization, feature scaling, and variable weighting to refine model accuracy. Once validated, the model will be expanded to enable clinicians to upload new de-identified patient data and automatically generate severity classifications based on the trained UC model. The platform will also allow users to visualize model outputs and identify which biomarkers and clinical variables most influence predictions, offering interpretability alongside performance. These updates will allow the system to be evaluated under real-world clinical conditions, improving its reliability and practical utility. Together, these advancements will transition the platform from a research prototype to a validated clinical decision-support tool, demonstrating how machine learning can be responsibly deployed to enhance predictive precision medicine and improve UC management.

■ Acknowledgments

This work was supported in part by the William Welles Bouton Award, which provided a $350 grant. I am deeply grateful to Dr. Soula Koniaris (Pediatric Gastroenterology, RWJBH Medical Group) for her guidance on the clinical relevance of IBD biomarkers, Dr. Sree Mallikarjun (UVA Data Science; Chief Data Scientist & Head of AI Innovation at Octus) for his mentorship and insights on model development and methodological planning in the ML framework, and to Dr. Sarag Boukhar (Pathology, RWJBH Medical Group) for his insights into the pathology of UC. I would also like to thank Dr. Rosa-

DOI: 10.36838/IJHSR87.65

lind Abbott (Carnegie Mellon University) for her mentorship and project advisement. Their expertise, support, and encouragement were invaluable in shaping this project.

References

1. Lynch, W. D.; Hsu, R. Ulcerative Colitis. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2025.
2. *Diagnosis of Ulcerative Colitis - NIDDK*. National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/digestive-diseases/ulcerative-colitis/diagnosis (accessed 2025-07-27).
3. *Definition & Facts of Ulcerative Colitis - NIDDK*. National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/digestive-diseases/ulcerative-colitis/definition-facts (accessed 2025-08-27).
4. Kulkarni, C.; Liu, D.; Fardeen, T.; Dickson, E. R.; Jang, H.; Sinha, S. R.; Gubatan, J. Artificial Intelligence and Machine Learning Technologies in Ulcerative Colitis. *Ther. Adv. Gastroenterol.* **2024**, *17*, 17562848241272001. https://doi.org/10.1177/17562848241272001.
5. Mallikarjun, S.; Abbasi, A.; University of Virginia, McIntire School of Commerce. Big Data Projects. In *Quantitive Methods*; CFA Institute, 2020; pp 518–590.
6. Khor, B.; Gardet, A.; Xavier, R. J. Genetics and Pathogenesis of Inflammatory Bowel Disease. *Nature* **2011**, *474* (7351), 307–317. https://doi.org/10.1038/nature10209.
7. Porter, R. J.; Kalla, R.; Ho, G.-T. Ulcerative Colitis: Recent Advances in the Understanding of Disease Pathogenesis. *F1000Research* **2020**, *9*, F1000 Faculty Rev-294. https://doi.org/10.12688/f1000research.20805.1.
8. Singh, S.; Ananthakrishnan, A. N.; Nguyen, N. H.; Cohen, B. L.; Velayos, F. S.; Weiss, J. M.; Sultan, S.; Siddique, S. M.; Adler, J.; Chachu, K. A. AGA Clinical Practice Guideline on the Role of Biomarkers for the Management of Ulcerative Colitis. *Gastroenterology* **2023**, *164* (3), 344–372. https://doi.org/10.1053/j.gastro.2022.12.007.
9. *What is IBD Plexus? | Crohn's & Colitis Foundation*. https://www.crohnscolitisfoundation.org/research/plexus/what-is-ibd-plexus (accessed 2025-11-06).
10. Amazon Machine Learning - Developer Guide.
11. Gulati, J. *Logistic vs SVM vs Random Forest: Which One Wins for Small Datasets?*. MachineLearningMastery.com. https://machinelearningmastery.com/logistic-vs-svm-vs-random-forest-which-one-wins-for-small-datasets/ (accessed 2025-11-06).
12. Rodríguez-Pérez, R.; Bajorath, J. Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. *J. Comput. Aided Mol. Des.* **2022**, *36* (5), 355–362. https://doi.org/10.1007/s10822-022-00442-9.
13. Maconi, G.; Bolzoni, E.; Giussani, A.; Friedman, A. B.; Duca, P. Accuracy and Cost of Diagnostic Strategies for Patients with Suspected Crohn's Disease. *J. Crohns Colitis* **2014**, *8* (12), 1684–1692. https://doi.org/10.1016/j.crohns.2014.08.005.
14. Strimbu, K.; Tavel, J. A. What Are Biomarkers? *Curr. Opin. HIV AIDS* **2010**, *5* (6), 463–466. https://doi.org/10.1097/COH.0b013e32833ed177.
15. Alghoul, Z.; Yang, C.; Merlin, D. The Current Status of Molecular Biomarkers for Inflammatory Bowel Disease. *Biomedicines* **2022**, *10* (7), 1492. https://doi.org/10.3390/biomedicines10071492.
16. Zamani, H.; Barzin, G.; Yousefinia, M.; Mohammadkhani, A.; Ostovaneh, M. R.; Sharifi, A. H.; Tayebi, S.; Malekzadeh, R.; Ansari, R. Diagnostic Value of Fecal Calprotectin in Patient with Ulcerative Colitis. *Middle East J. Dig. Dis.* **2013**, *5* (2), 76–80.
17. Walker, T. R.; Land, M. L.; Kartashov, A.; Saslowsky, T. M.; Lyerly, D. M.; Boone, J. H.; Rufo, P. A. Fecal Lactoferrin Is a Sensitive and Specific Marker of Disease Activity in Children and Young Adults with Inflammatory Bowel Disease. *J. Pediatr. Gastroenterol. Nutr.* **2007**, *44* (4), 414–422. https://doi.org/10.1097/MPG.0b013e3180308d8e.
18. *Erythrocyte Sedimentation Rate (ESR): MedlinePlus Medical Test*. https://medlineplus.gov/lab-tests/erythrocyte-sedimentation-rate-esr/ (accessed 2025-08-28).
19. He, T.; Wang, K.; Zhao, P.; Zhu, G.; Yin, X.; Zhang, Y.; Zhang, Z.; Zhao, K.; Wang, Z.; Wang, K. Integrative Computational Approach Identifies Immune-relevant Biomarkers in Ulcerative Colitis. *FEBS Open Bio* **2022**, *12* (2), 500–515. https://doi.org/10.1002/2211-5463.13357.
20. *Support Vector Machine (SVM) Algorithm*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/support-vector-machine-algorithm/ (accessed 2025-07-27).
21. Saxena, V. Support Vector Machine and Its Applications in Information Processing. Thesis, Massachusetts Institute of Technology, 2004. https://dspace.mit.edu/handle/1721.1/29404 (accessed 2025-07-27).
22. *1.4. Support Vector Machines*. scikit-learn. https://scikit-learn/stable/modules/svm.html (accessed 2025-07-27).
23. *Drug Classification*. https://www.kaggle.com/datasets/prathamtripathi/drug-classification (accessed 2025-07-27).

Author

Aayushi Saxena is a senior at The Lawrenceville School. She is passionate about data science, humanities, and policy. Specifically, she wants to explore how technology innovation can be democratized and implemented to streamline convoluted systems (e.g., healthcare and law), improving access and creating meaningful impacts in people's everyday lives.