SMU Data Science Review

Volume 5 | Number 2 Article 9

2021

Machine Learning Approach to Distinguish Ulcerative Colitis and Crohn's Disease Using SMOTE (Synthetic Minority Oversampling Technique) Methods

Kris Ghimire
Southern Methodist University, kghimire@mail.smu.edu

Walter Lai
Southern Methodist University, wlai@smu.edu

Yasser Omar
Columbus University, dr_yaser_omar@yahoo.com

Thad Schwebke
Southern Methodist University, tschwebke@mail.smu.edu

Jamie Vo
Southern Methodist University, jamiev@mail.smu.edu

Recommended Citation

Ghimire, Kris; Lai, Walter; Omar, Yasser; Schwebke, Thad; and Vo, Jamie (2021) "Machine Learning Approach to Distinguish Ulcerative Colitis and Crohn's Disease Using SMOTE (Synthetic Minority Oversampling Technique) Methods," *SMU Data Science Review*: Vol. 5: No. 2, Article 9.
Available at: https://scholar.smu.edu/datasciencereview/vol5/iss2/9

1. Introduction

The human digestive system is a complex ecosystem inhabited by hundreds of intestinal microflorae. Improper interaction between gut microbiota and the mucosal immune system can result in chronic inflammation of the gastrointestinal (GI) tract, also known as Inflammatory Bowel Disease (IBD). Researchers have disputes on whether the disease is considered an autoimmune disease, but it has been correctly categorized as the autoimmune system attacking harmless beings in the gastronomical tract. IBD may cause unexplained vomiting, persistent pain, and/or anemia. In severe cases of IBD, patients who are in late stages may have permanent damage including merging of organs (e.g., intestines to kidneys), inability to control secretion of waste, or requiring external methods to rid of body toxins (Inflammatory bowel disease (IBD), n.d.).

Two major forms of IBD have steadily increased globally. Ulcerative Colitis (UC) primarily affects the lining of the large intestine, and Crohn's Disease (CD) affects all layers from the mouth to the anus, leaving various healthy parts in between the inflamed areas. One other form of IBD consists of IBD without differentiating features, called indeterminate colitis (What is IBD?, n.d.)

Commonalities between UC and CD include when the disease presents itself. The onset begins during adolescence or early adulthood, although it can occur at any point in a person's life. Neither of the diseases discriminates against genders and are equally prevalent. A significant concern with the two conditions is the similarities in symptoms, resulting in difficulties in differentiating between the two. Due to the two diseases affecting different areas of the human body, an understanding of which inflammation affects the patient is critical in treatment (Cherney, 2020). While both diseases affect the large intestine, the differences in areas allow physicians to treat the related diseases properly.

A difference between UC and CD is that UC affects an entire area continuously. UC will be constant from the start of the affected areas to where the inflammation ends, unlike CD, which may have breaks in the inflicted regions. UC is strictly contained in the large intestine, while CD can occur at any section between the mouth and anus. Another significant difference is that UC primarily affects the colon's outer lining, contrasted by CD, affecting any layer (Cherney, 2020). The inflammation is distinct due to the disease's areas, such as the lining layers to the area in the digestive process. As a result of CD's effect throughout the gastronomical tract, many other systems can be affected, such as the eyes or liver (What is IBD?, n.d.).

Despite the prevalence of IBD, the cause is not yet widely understood. According to the Crohn's & Colitis Foundation of America, as many as 70,000 new IBD cases are diagnosed each year in the US; however, the disease's exact cause has not been found. Research suggests that patients inherit genes that make them susceptible to IBD.

Nevertheless, the primary gene producing the secretive protein linked to IBD has not been discovered. Even though there are breakthrough therapies, such as Monoclonal anti-TNF antibodies, only 30% of patients with IBD show a response to the drug. Among those that respond to treatment, a significant percentage of patients will relapse over time. Therefore, there is a therapeutic need to develop novel agents that target diverse molecular pathways involved in the pathophysiology of IBD to supply the proper medical treatments (Data and Statistics, 2020).

Like many other diseases, UC and CD can progress in stages. While the disease may be deceiving because it can come in waves, with periods in which little to no symptoms appear, the disease itself is progressing. As the patient ages and the disease continue without treatment, the likelihood of irreversible effects increases dramatically. Patients with CD who allow the disease to continue without seeking treatment may cause them to be more susceptible to colon cancer. If the disease is diagnosed in the initial stages, there may be methods to prevent specific symptoms or biological degradation from occurring (Cherney, 2020).

Classifying UC severity stages and disease type allows physicians to take proper steps in earlier phases and supply proper medical therapy. By diagnosing the disease before its progression, the prevention of surgery, irreversible damage, and decline in quality of life can be achieved. Physicians currently differentiate disease stages based on symptoms and lab tests such as hemoglobin (Hgb), C-Relative Protein (CRP), Erythrocyte Sedimentation Rate (ESR), fecal calprotectin tests, serum albumin, imaging, and endoscopic procedures, which are not always capable of finding the primary disorder. The researcher in this paper focuses on building machine learning (ML) approaches to differentiate the diseases using endoscopic and histological data.

Endoscopy plays a significant role in the diagnostic workup for patients who are suspected of having IBD. The procedure consists of inserting a long thin tube to observe the gastrointestinal mucosa and the biopsy of different sites. Therefore, endoscopy is a macroscopic investigation of IBD, primarily aimed at detecting neoplastic lesions early and preventing colorectal cancer in high-risk patients. Endoscopic features significant in training ML algorithms include mucin depletion, epithelial changes, cryptitis polymorphs, segmental colitis/enteritis, etc.

Histology is the microscopic examination of biopsy images taken from multiple upper and lower GI segments. By applying a microscopic analysis to the tissues and cells of the biopsy, clinicians can observe changes in cell formation, clustering of T or B cells in response to inflammation, and/or crypt branching. Examples of histological features are crypt architectural distortion, basal plasmacytosis, and neutrophilic activity.

A significant attribute studied to find IDB's presence includes crypt architecture and lamina propria cellularity, observable in Figure 1 and Figure 2 with varying distortions. The crypt architecture of a healthy colon is singular, with a single crypt that reaches deep into the tissue. In the presence of diseases such as IBD, distortion can occur where the crypts become branched or lost, resulting in improper absorption and malnutrition. Other changes in the architecture include the presence of an abscess, abscess formations, and crypt inflammation (Crypt distortion, n.d.).

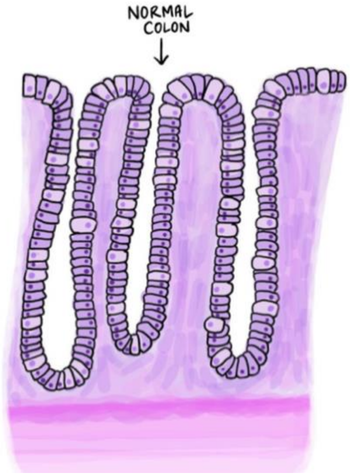

Figure 1: Normal Colon (Crypt distortion, n.d.)

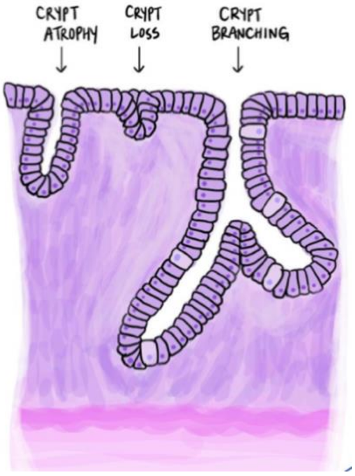

Figure 2: Colon with varying distortions (Crypt distortion, n.d.)

Lamina propria cellularity is the thin mucus layers line the body's inner tubing (e.g., respiratory tract). In the presence of IBD, inflammation, missing areas of lamina propria, or the presence of lamina propria granulomas are significant indicators of the disease. The layer has important immunological cells, which can result in clustering/increase in lymphocytes cells.

Due to the many similarities between CD and UC, it is dramatically tricky for physicians to differentiate between them. Since discovering the disease early on is significant in preventing progression, there is a strong need for a quick and accurate diagnosis and distinguish between the two diseases. Applying machine learning methods to the issue can reduce the potential for human error and incorrect classification.

The research discussed below contains machine learning models, such as random forest and support vector machines, to analyze histological and endoscopic data and differentiate between UC and CD and UC, CD and normal. By applying ML, physicians can quickly and accurately screen many patients for IBD. In addition, patients can prevent the progression of the disease since a determination of UC, Chron’s, or normal will be made.

2. Background

The following section provides a background on the classification algorithms and evaluation metrics used.

Many machine learning algorithms will be applied, including support vector machine (SVM), decision trees, neural networks (NN), K Nearest Neighbors, Logistic Regression, and Ada Boost. SVM and decision trees are supervised machine learning methods where the outcome is supplied for the model to determine the accuracy of its algorithm. On the other hand, neural networks allow the algorithm to find patterns in the data despite not understanding the outcome. K Nearest Neighbors classifies samples by finding the most similar samples (neighbors) and picking the target class that is most represented in those neighbors (Brownlee, Boosting and AdaBoost for Machine Learning, 2016; Brownlee, K-Nearest Neighbors for Machine Learning, 2016). Logistic Regression samples by assigning coefficients to variables and discovering the linear relationship between the variables with the log odds of the target class. The log odds are “the probability of the event divided by the probability of not the event” (Brownlee, Boosting and AdaBoost for Machine Learning, 2016; Brownlee, Logistic Regression for Machine Learning, 2016). Ada Boost classifies samples by sequentially creating a classifier, then building a second classifier that “attempts to correct the errors from the first model,” then adding classifiers on the previous (Brownlee, Boosting and AdaBoost for Machine Learning, 2016). A more in-depth explanation of the algorithm's implementations follows.

Observable in Figure 3, support vector machines strive to create as much distance between classes as possible while maintaining the inclusion of data points in the same class. Unlike other classification algorithms that only use data points, SVM brings the data to higher dimensions where the separations are distinct. The main concern with SVM is that the data transformation is irreversible, unlike data manipulation such as applying a logarithmic transformation.

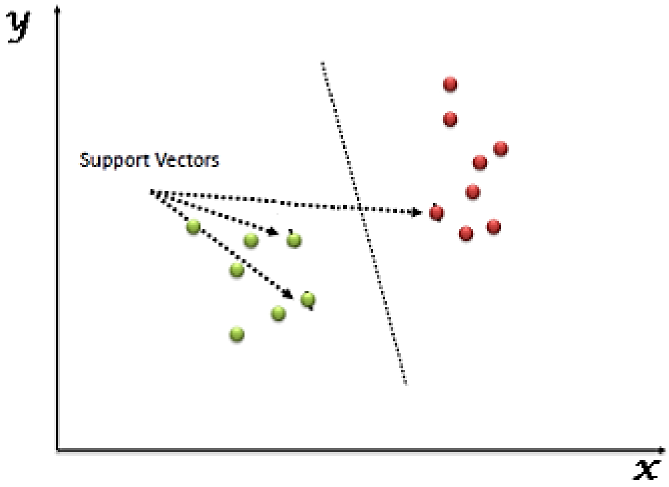

Figure 3: Example SVM (Ray, 2017)

Decision trees are a method to classify data into segments using a yes/no logic. An example of this may be a generic question; if yes, a second question further separates the data; if no, the sample is classified into class A. Visible in Figure 4, the class decides the class using a series of questions, typically based on attributes where the similarities/patterns are determined.

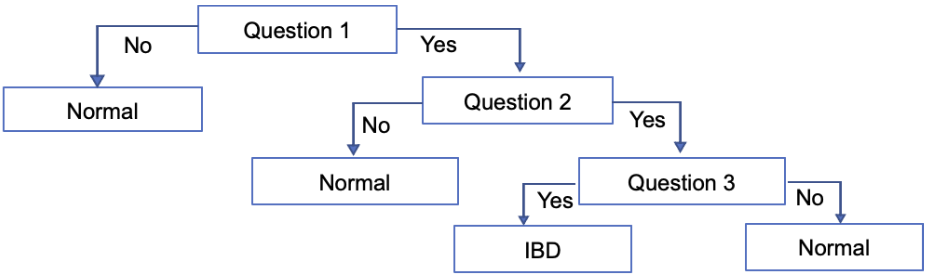

Figure 4: Example Decision Tree

Lastly, neural networks (NN) are modeled after human brain activity, indicative of the name. NN has a multitude of layers where transformations are performed. The data is loaded into the first layer, called an input layer, after which all layers between the acceptance and exit are "hidden" layers. The hidden layers contain nodes with weights associated. The method uses backpropagation, where segments of the data are loaded and the prediction is determined. Weights are then sent back to the proper nodes and adjusted as needed. As each iteration progresses, the tuning of the weights continues until a global error minimum is reached. The output is the classification, disease types, and disease stages. In Figure 5 below, x1 and x2 are input features; the red box represents weights, the green box represents the bias term, the yellow box is the activation function of distinct types, and y is an output class.

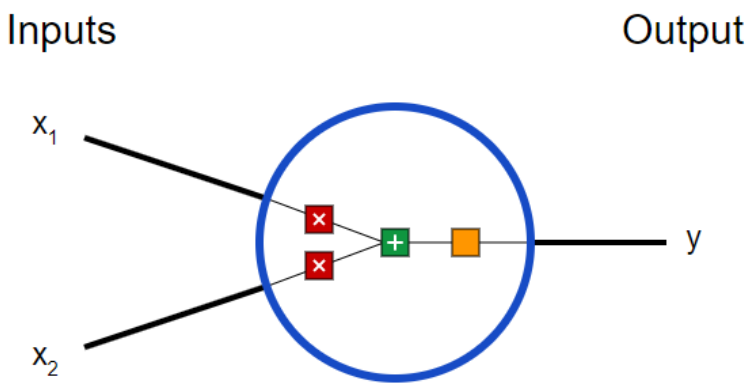

Figure 5: Example Neural Network (Zhou, 2019)

The metrics used to follow many of those discussed in the literature, located in the related works section, including accuracy, precision, recall, and area under the curve.

For supervised learning, evaluating the model on unseen data will improve the model's generalizability to predict irritable bowel disease stages for new cases. Accuracy is calculated as the sum of true positives and negatives over the total sum of a correct and incorrect prediction. Accuracy can be useful for this research because accurate predictions of disease type and disease stages.

Precision is defined as the number of predictions made that are correct out of all the positive predictions (Brownlee, How to Use ROC Curves and Precision-Recall Curves for Classification in Python, 2018). It signifies the model's ability to minimize false positives (Brownlee, How to Use ROC Curves and Precision-Recall Curves for Classification in Python, 2018). A recall is defined as the number of true positives out of all the positive samples in the dataset (Brownlee, How to Use ROC Curves and

Precision-Recall Curves for Classification in Python, 2018). A high recall indicates a low false-negative rate, being especially valuable in healthcare since failing to diagnose a diseased patient will delay treatment and potentially cause harm.

F1 is “a way of combining the precision and recall” (Wood, n.d.). It is the product of precision and recall, divided by the sum of the two (Wood, n.d.). F1 is the statistic used to optimize the grid search when tuning the classifier models. This ensures that resulting models will not improve precision at the expense of recall or vice versa.

The area under the curve (AUC)-Receiver Operating Characteristic (ROC), better known as the ROC Curve, is an excellent method for measuring the performance of a classification model. The True Positive Rate (TPR) is plotted against the False Positive Rate (FPR) for the probabilities of the classifier predictions. Then, the area under the plot is calculated. The greater the area under the curve, the better the model is at distinguishing between classes. AUC curves will be analyzed alongside the ROC (receiver operator curve) to find the model’s power (Brownlee, How to Use ROC Curves and Precision-Recall Curves for Classification in Python, 2018). The metric Kappa evaluates the accuracy of the model compared to random guesses (Widmann, 2020). A large value of Kappa means that the model’s predictions agree with the true distribution of the target class and controls class imbalance (Widmann, 2020).

2.1. Machine Learning

This section will survey studies related to machine learning for diagnosing IBD and discuss the research gap. This study determines performance by applying general machine learning algorithms to an IBD dataset and comparing the model to a human pathologist’s diagnosis.

The research performed in this study is related to four other studies that used the same data set but with different modeling techniques. The data set shared by these studies is based on Dr. Simon S. Cross’ study in 1999. Cross’ study used 809 endoscopic colorectal biopsies (Cross S. S., 1999), in addition to the histological data for the patients of the biopsies. From the complete set, 165 of the biopsies showed no signs of IBD and 644 were confirmed as having IBD. Of the 644 IBD biopsies, 473 had UC, while 171 had CD. The data was cross partitioned into two relevant partitions: 1) the full set of 809 biopsies (all IBD and normal) and the whole IBD set consisting of 644 IBD confirmed biopsies. The first partition (all IBD and normal) was used to determine if a patient had IBD or did not. The second partition (all IBD) was used to find whether the patient had UC or CD.

Asmaa Mohamed Hassan and Yasser M. K. Omar (Hassan & Omar, 2020) used a systematic workflow to perform their study. The workflow starts with loading the whole IBD and normal data set. Integer and one hot encoding were then used to convert the categorical variables to binary vectors. Five-fold cross-validation was used, and each fold was split into two subsets, training 80% of the data and test 20% of the data. After the preprocessing of the data was complete, two different approaches were used. The first used SVM and an Artificial Neural Network (ANN) to classify the diagnosis between UC and CD. The second approach classified the diagnosis between normal, active CD, inactive CD, active UC, and inactive UC. Accuracy, recall, and specificity were used as the performance metrics. Their best model was the ANN with 82.7% accuracy, 90% recall, and 54.4% specificity.

Cruz-Ramirez et al. compared different feature selection techniques to classify the three diseases. (Cruz-Ramírez, Acosta-Mesa, Barrientos-Martínez, & Nava-Fernández, 2006). They compared a Naïve Bayes model that used all features to other more sophisticated feature selection algorithms and the results from an experienced pathologist. In their first experiment, Naïve Bayes with all features outperformed the other models in sensitivity (96%) and specificity (69%) when distinguishing normal from any diseased state: CD or IBD. In their second experiment, Naïve Bayes barely outperformed the other feature selection algorithms in sensitivity (65%) and specificity (68%), showing the difficulty in selecting good histological features between CD and IBD. This paper is useful in that it shows Bayesian networks can explain interactions between certain histological features. This research plans to incorporate Bayesian networks to help facilitate communication and cooperation with domain experts. Additionally, Cruz- Ramirez et al. could not produce models that outperformed the experienced pathologist, suggesting using image data to augment structured tabular data.

Cross' study used Receiver Operating Characteristic (ROC) Curve Analysis to derive accuracy indexes, specifically the area under the curve (AUC), to interpret UC or CD classification compared to normal patients. McNemar's tests were then used to compare the AUC. The advantage of using the ROC curve is finding the optimal thresholds. Once an optimal point was reached, sensitivity, specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), and the Kappa statistic were calculated with a 95% confidence interval. The results of Cross' studies produced a sensitivity for the diagnosis of CD or UC in the range of 40% to 82% and a specificity in the range of 73% to 98%. The conclusion suggested that there was room for decision-support systems in the histopathological diagnosis of IBD to improve sensitivity and PPV (Cross S. S., 1999).

Many other sources were used to understand the different machine learning techniques used. Boyd et al. (Boyd, et al., 2018) used CAGE (Cap Analysis of Gene Expression) data on biopsy samples from the descending colon (a portion of the large intestine) (taken from 94 IBD patients and a control group) to distinguish between active UC and CD. They were able to classify the diseases with 95% accuracy using random forest. Lichy Han et al. (Han, et al., 2008) used a novel pathway-based approach called PROBS (probabilistic pathways score) to calculate individualized pathway scores for the same classification. Their method resulted in an aggregated AUC of 0.82 using a random forest algorithm.

Khorasani et al. (Khorasani, Usefi, & Peña-Castillo , 2020) used the feature selection algorithms DRPT and SVM to produce a model that classifies between healthy and UC patients based on the expression values of thirty-two colon genes. They used precision metrics to evaluate their model and previously developed models. To increase the model's robustness, they combined several independent gene expression datasets with the NCBI Gene Expression Omnibus (GEO) database. They discovered the thirty-two most relevant genes in classifying UC patients, some of which include FAM118A, LIPF, MMP2, DMTN, PPP1CB, etc. Their model's prediction performance increased precision over the BioDiscML model (a biomarker discovery software model). The researchers in this paper will be using feature selection and SVM methods to enhance the paper results and increase robustness.

Mossotto et al. (Mossotto, et al., 2017) presented a machine learning approach to disease distinction like the research discussed in the methods section. Mossotto's data consisted of 287 pediatric patients, including UC, CD, and unclassified disease patients. Unsupervised clustering algorithms and hierarchical clustering were used to visualize the relationship between patients and their clinical characteristics. The supervised machine learning model, SVM, was run using linear kernel on three diverse types of datasets (endoscopic, histological, and joint endoscopic/histological). The combined model outperformed with an accuracy of 82.7%, precision of 0.91, recall of 0.83, and F1-score of 0.87. Like Hassan et al. (Hassan & Omar, 2020), the SVM model will be used, and methods by Mossotto et al. (Mossotto, et al., 2017) will be used in conjunction to increase accuracy.

Smillie et al. (Smillie, et al., 2019) used single-cell RNA sequencing (scRNA-seq) on colon biopsy specimens. They showed significant cell proportions and gene expression in healthy stages of non-inflamed and inflamed patients. Risk genes were nominated across loci, cells, and their functions were predicted and assembled into core pathways, resulting in a controlling mechanism of the disease. Some techniques implemented include clustering cells into epithelial, stromal, and immune compartments, principal component analysis (PCA) for feature reduction, t-SNE visualizations, and k-NN graphs. Random Forest was trained with 1,000 trees and default parameters to produce an out-of-bag error of 10.7%. Again, the researchers will be utilizing PCA to reduce the dimensionality of the datasets.

Other research studies were based on functions in the digestive system. The expectation was it would result in more accurate predictions. D'haeseleer (D'haeseleer, 2005) supplied an overview of gene expression clustering, saying, "How does gene expression clustering work?". The paper recommends evaluating gene clustering by reviewing the genes' function in the same cluster. The clustering algorithm is effective if two genes are in the same cluster, share similar functions, contribute to the same pathways, and have common protein-protein interactions. An extensive EDA (Exploratory Data Analysis) was performed on the resulting clusters to find the relationships between genes to cluster individual IBD genes into broader categories. Other researchers in the field may then use the resulting clusters to discover biological relationships and reduce dimensionality.

Clustering algorithms called generative models have been developed and showed promising gene expression clustering results. Generative models are weakly supervised; records are labeled or clustered according to multiple, different external datasets. Dutta et al. (Dutta, Saha, Pai, & Kumar, 2020) applied weakly supervised learning and generative models to the task of gene clustering in "A Protein Interaction Information Based Generative Model for Enhancing Gene Clustering." This contribution was novel since Dutta and colleagues used protein-protein interaction and Gene Ontology Consortium data to cluster genes. The clustering algorithm achieved a 0.941 Silhouette and 0.994 Biological Stability Index, outperforming the current gene clustering algorithms at the time. The methods discussed above are applied to the genes specific to IBD, reducing dimensionality and unveiling the dependencies between certain genes.

New gene clustering packages were applied that buckets genes into broader categories based on distinctive features. Yoon et al. (Yoon, et al., 2019) developed an R Package called GSCluster that clusters genes into "gene-sets" based on how often the genes are expressed together. This is a novel contribution in that the Yoon combined gene-set clustering with weights based on protein-protein interactions. Jeggari et al. (Jeggari & Alexeyenko, 2017) developed an R package called NEArender, which clusters individual genes into broader categories based on the pathways the genes contribute to. The study used network enrichment analysis to cluster genes and is a novel contribution due to parametric methods to estimate error instead of prior non-parametric methods. This increases the clustering speed.

Manandhar et al. (Manandhar, et al., 2021) used gut microbiome data for a disease classification. Using 331 CD and 141 UC samples, they found 117 differential bacterial taxa (LEfSe: LDA > 3). They hypothesized that diagnostic classification of IBD can be achieved by using machine learning models. To meet their study goal, they collected 16S rRNA metagenomics data from the American Gut Project. Researchers used five different algorithms, random forest (RF), decision tree (DT), elastic net (EN), support vector machine with the radial kernel (SVM), and neural networks (NN), to make classifications. The disease type (CD and UC) was classified with 83% accuracy. Other metrics they used to validate their model include area under the receiver operating characteristic curves (AUC), sensitivity, specificity, precision, and F1.

Mahapatra and colleagues developed a shape asymmetry measure based on image intensity values. (Mahapatra, Schueffler, Tielbeek, Buhmann, & Vos, 2012). They collected MRI (Magnetic Resonance Imaging) images from twenty-six patients with known diagnoses of normal or IBD. They used SVM, random forest, and a Bayesian classifier with their newly developed feature to achieve sensitivities and specificities that outperformed classifiers that used current image transformation techniques. This paper provides multiple methods of extracting features from biopsy images for dimensionality reduction.

The studies by Cross et al., Hassan et al., Cruz-Ramirez et al., and Tan et al. all worked on the same dataset by Cross et al.. Still, each used a small selection of classification algorithms with their cross-validation strategy and feature selection. This research uses the same dataset with a consistent cross-validation strategy to evaluate the performance of a wide range of classifiers and feature sets. This will help determine the best classification algorithm and set of features to use for diagnosing IBD.

3. Machine Learning Pipeline

The research consists of building a classifier to discriminate between UC and CD and another classifier to distinguish between normal, UC, and CD. Figure 6 describes the workflow used. The Data Cleaning component was composed of upper casing and removing leading and trailing spaces from all text.

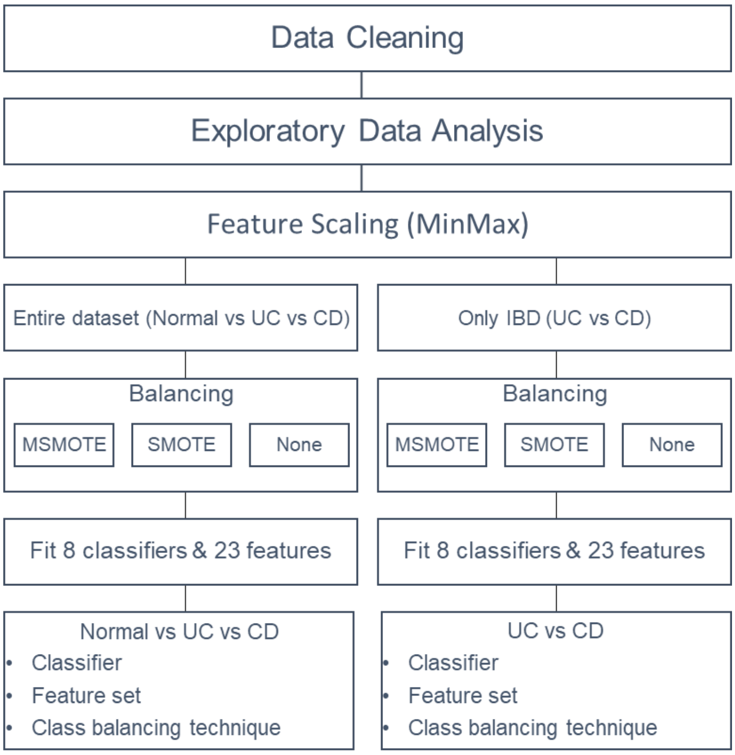

Figure 6: Model Pipeline

The Exploratory Data Analysis component in Figure 6 consisted of checking normality and feature dependence. The Feature Scaling component entailed scaling all the features relative to the minimum and maximum of the column. This resulted in a cleansed dataset.

Six datasets were created from the cleaned dataset. This represents the fourth and fifth layer from the top of Figure 6:

1. Cleaned dataset with only Diseased Patients (UC or CD)
2. Cleaned dataset with only Diseased Patients with SMOTE to generate new samples for CD.
3. Cleaned dataset with only Diseased Patients with MSMOTE to generate new samples for CD.
4. Cleaned dataset.
5. Cleaned dataset with SMOTE to generate new samples from the minority classes (CD and normal)
6. Cleaned dataset with MSMOTE to generate new samples from the minority classes (CD and normal)

Datasets 1, 2, 3 were used to classify between CD and UC. Datasets 4,5,6 were used to classify between normal, CD, and UC. Eight classification algorithms were selected. For each algorithm, a grid of parameter values was created for parameter tuning.

The twenty-three independent variables were ranked from most correlated to least against the response variable. Twenty-three sets of features were created. The first set contained the single, most correlated, independent variable. The second set contained the top two independent variables. The third set contained the top three independent variables and so on. The twenty-third set contained all the independent variables.

In the sixth layer from the top of Figure 6, eight classification algorithms were tuned on all twenty-three sets of features for each of the six datasets. The classification algorithm and feature set with the best average test F1 score for each dataset were selected. In the bottom layer of Figure 6, the best model for datasets 1, 2, and 3 is presented in the 'Results' section as the best model for Classification Task 1: CD versus UC. The best model built for datasets 4, 5, 6 is presented as the best model for Classification Task 2: normal versus CD versus UC.

The flowchart from Figure 6 was inspired by prior work (Hassan & Omar, 2020). The model was evaluated by averaging the five different test set F1 values for each cross-fold iteration. The description of 5-fold cross-validation was borrowed from an online source (Brownlee, Boosting and AdaBoost for Machine Learning, 2016; Brownlee, Gentle Introduction to k-fold cross validation, 2018)

4. Data Summary

The dataset referred to in this paper is sourced from the Department of Histopathology, Royal Hallamshire Hospital in Sheffield, United Kingdom (Cross S. S., 1999). The dataset has 809 endoscopic colorectal biopsies, of which 165 are normal (control group), 473 are UC, and 171 are CD (paraphrased from Hassan and Yasser's earlier analysis (Hassan & Omar, 2020)). For each biopsy, a single experienced observer recorded histological features such as the presence of inflammation or the prevalence of cell types (such as polymorphs) (Cross S. S., 1999). Other demographic data recorded include the age and sex of the patients. This led to a dataset consisting of twenty-three independent variables, primarily focused on features concerning the crypt architecture and lamina propria cellularity. Each biopsy also had the response variable, which was the patient's diagnosis confirmed through endoscopy, radiology, and microbiology (Cross S. S., 1999). While there was added diagnosis (initial and observing pathological), the features were removed prior to applying the ML algorithms to mimic a practitioner's initial appointment with the patient.

4.1. Data Preparation

Most of the data cleansing consisted of user input error, which results in a higher-class indication than the three diagnoses of normal, CD, and UC. In addition, the dataset received is in an ordinal format. More transformations were performed to convert the data to binary classifications and the original test results for interpretation.

This research uses a dataset that consists of primarily categorical variables (Cross S. S., 1999). The features are dummy coded where each categorical variable's levels are converted into dichotomous, binary variables. The decision to manage categorical variables in this manner was influenced by Dr. Yasser's work on the same dataset and allowed ML digestion of the data (Hassan & Omar, 2020).

Earlier work on the current dataset also addressed issues related to scaling (Hassan & Omar, 2020). Due to few continuous variables existing in the dataset, those which exist are on varying scales. Each continuous variable is scaled to their respective averages, fixed at zero, where the values will be converted to the number of standard deviations away from the mean. The data itself did not have any missing or duplicate observations.

The dataset has an imbalanced representation of the diagnosis class. SMOTE is used to generate more samples of the minority classes to meet the same ratio as the majority by using K Nearest Neighbors to produce additional samples. The advantage of SMOTE is that all the data is kept as opposed to down-sampling, where the data is discarded (Li & Lu, 2019).

Another form of class balancing imposed is Modified Smote (MSMOTE). Modified smote also generates new samples for the minority class but differs from SMOTE in producing new samples. SMOTE selects minority samples by randomly choosing one of its neighbors to make a new sample. MSMOTE is more selective in which neighbors were chosen to construct the new sample (Hu, Liang, Ma, & He, 2009). When MSMOTE sees a minority sample with all K-neighbors that are minority samples, it will act similarly to SMOTE by picking an indiscriminate neighbor from which to generate a new sample. If MSMOTE receives a minority sample where all the neighbors are not minority samples, it will pick the nearest neighbor to generate a new sample (Hu, Liang, Ma, & He, 2009).

Prior studies created models that either discriminating between UC versus CD and/or between normal versus CD versus UC. This research performed both classification tasks. Cross-validation techniques are applied to improve the generalizability of other researcher's models (Cross & Harrison, 2002). They split the first 540 patients as the train set and preserve the remaining 269 patients as the test set. The authors of this paper use five-fold cross-validation instead, as it offers a more accurate measure of generalizability than a traditional train-test split.

5. Results

A comparison of the performance of algorithms and class balancing techniques follows. Note that random forest was the best performing algorithm in both classification tasks, and MSMOTE was the best class balancing technique in both classification tasks.

5.1. UC versus CD

The best model, seen in Figure 7 and Figure 8, for discerning UC from CD, was a random forest Classifier with an average test fold F1 of 0.989 over five-fold cross-validations on all the patients that had UC or CD. This model had a recall and accuracy of 0.988. The data preparation for this model was MSMOTE.

This paper's performance on classifying CD versus UC was compared to other models in the literature. Cross et al. 1999 reported the best sensitivity observed in the literature that ranged between 40% to 82% and a specificity of 73% to 98% (Cross S. S., 1999). This paper's best model, random forest, for CD versus UC reported an accuracy of 98.8%, a precision of 98.8%, and a sensitivity of 98.8%. This paper's model outperformed the current best model for diagnosing CD versus IBD disease, given that the patient already has IBD by an increase of 16.8% to 58.8% in sensitivity while maintaining high precision.

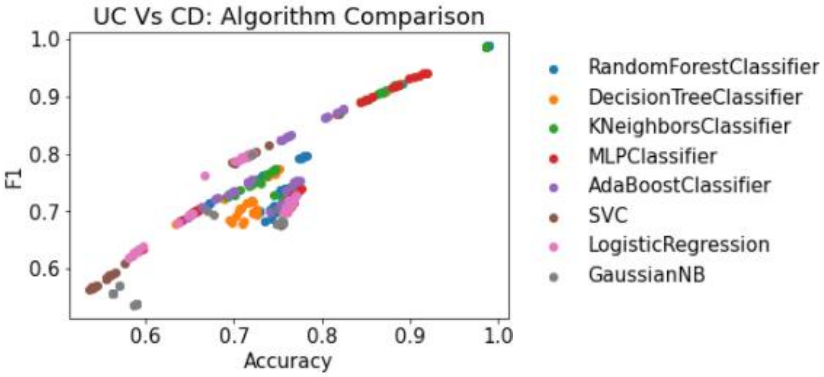

Figure 7: Algorithm Performance of CD versus UC

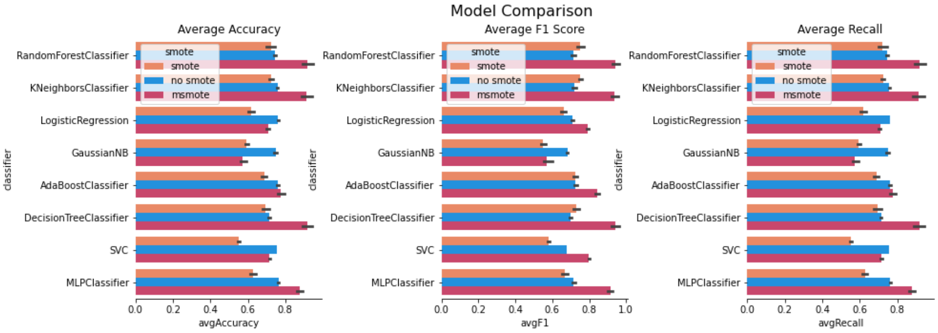

Figure 8: Each Algorithm’s Best Model metrics for CD vs UC.

5.2. UC versus CD versus Normal

The best classifier, seen in Figure 9 and Figure 10 for discerning between normal versus UC versus CD was a random forest Classifier with an average test fold F1 of 0.747 across five-fold cross-validations using the entire dataset. The average test fold recall and accuracy of this model was 0.729. The data preparation for this model was MSMOTE.

This model’s performance for classifying normal versus IBD as compared to other models in the literature. Cross et al. 1999 reported an accuracy of 88.75%, a sensitivity of 79-86%, and a specificity of 98-99% (Cross S. S., 1999). Tan et al. 2006 reported accuracy of 65.56%, the sensitivity of 77%, and specificity of 22.8% (Tan, Ng, & Erdogan, 2006). Cruz-Ramirez et al. reported sensitivity of 92-100% and specificity of

63-75% (Cruz-Ramírez, Acosta-Mesa, Barrientos-Martínez, & Nava-Fernández, 2006). Hassan et al. 2020 reported an accuracy of 82.7% and sensitivity of 79%, and specificity of 83.7% (Hassan & Omar, 2020). This paper's best model for normal versus CD versus UC reported an accuracy of 74.7%, a precision of 72.9%, and a sensitivity of 72.9%. Our model did not outperform the current best models. However, this paper's model's advantage over the current best model is that it discerns between all three states, normal, Crohn, and UC, with reasonable accuracy, recall, and precision compared to previous models, which only distinguish between normal and disease.

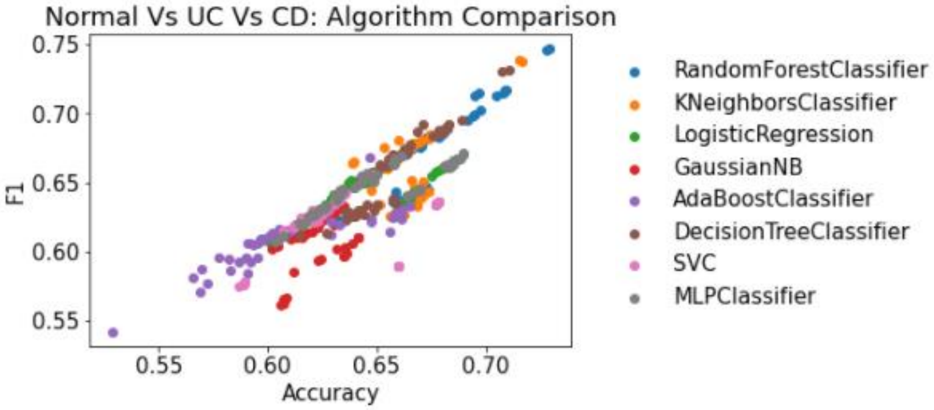

Figure 9: Algorithm Performance for CD vs UC vs Normal

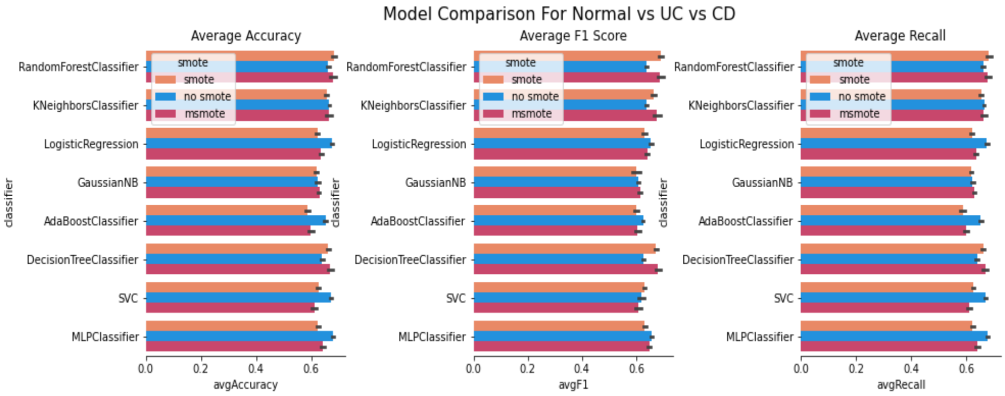

Figure 10: Each Algorithm's Best Model metrics for Normal vs CD vs UC.

6. Discussion

This paper produced a model that outperformed the literature's best model for classifying CD versus UC. This is due to better EDA, balancing, and other methods such as hyperparameter tuning. This paper did not produce a model that outperformed the literature's best model for classifying normal vs. IBD. While creating these models, the team discovered some limitations and opportunities for future work. Furthermore, a discussion of the impact and ethics of this research follows.

6.1 Limitations

The model was trained on a dataset that was not directly generated from biopsies. A trained pathologist had to derive histological features from patient biopsies (Cross S. S., 1999). This limits the generalizability of this paper's model. For this paper's model to predict a patient's diagnosis, both a biopsy and a trained pathologist must create the input features.

Furthermore, this paper's model requires all 23 of the features. This results in an overly complex model and may be susceptible to overfitting. Many of the features in this dataset may be correlated and redundant. Another limitation was the small size of the current data set. Our inference is limited to a particular group of patients.

Additionally, we did not have genomic data for the patients listed. The mapping of genes and changes in gene expression per disease type could provide more significant insight into differentiating the disease.

6.2 Future Research

Future research should reduce the number of features for predictive models and produce simpler, more interpretable models. Explainable models facilitate cooperation from subject matter experts, such as physicians and nurses, to have greater knowledge behind IBD.

Visible in Figure 8 and Figure 10, the diagnosis balancing method of MSMOTE consistently outperforms SMOTE despite not having class balancing. MSMOTE is sensitive to clusters in the dataset while SMOTE is not (Hu, Liang, Ma, & He, 2009). Thus, MSMOTE's superior performance may point to the existence of clusters of patients that are correlated with disease stages. Future work should consider identifying these clusters to reduce the complexity.

While there may be clusters in the dataset, outliers can be present. This research paper retained the full dataset, allowing for the possibility of particularly unusual patients to skew the model's predictions. A fascinating future investigation will identify clusters of patients based on symptoms and determine individual patients who do not fall within a designated grouping.

6.3 Impact

The impact of this research applies to practitioner care, pharmaceuticals, and patients who are affected by IBD. Better care and diagnosis can be decided by using the RF model since this model outperformed other models. Although the RF model effectively classifies IBD with high model metrics, areas of this work can be further enhanced by collecting current data. By distinguishing between the different types of IBD based on data and statistics, physicians can quickly and confidently decide treatment plans for the patient's current state.

An accurate model such as the ones presented in this paper can save the patient's life or reduce unnecessary costs. The model minimizes the possibility of misdiagnosing patients. For example, if a physician notifies a patient they are not inflicted with the disease when they are, the patient is at risk of colon cancer or intestinal damage (Cherney, 2020). According to the CDC, "The [average] hospitalization costs in 2014 were $11,345 for Crohn's disease and $13,412 for ulcerative colitis" (Data and Statistics, 2020). An accurate diagnostic model will allow the patient to receive and pay for the care they need.

7.4 Ethics

This study analyzed data from real-life patients. To comply with research ethics, the patients consented to have their health information be publicly available. Patient consent was confirmed in the "Methods" section of the study involving patient data. In addition, ensuring the patients understand the testing, how the data will be applied, and that their data is protected (Personally Identifiable Information (PII)) under the Privacy Act of 1974.

7. Conclusion

Using the power of data and ML algorithms, this study has successfully classified the IBD disease types and differentiated the diseased from normal patients. By applying multiple methods to distinguish between the two similar but complex and rare diseases, CD and UC, this paper has outperformed some of the previous work done on a similar dataset. Based on model evaluation metrics and other statistics, the RF was determined to be the best model. The analysis also shows that all twenty-three features are key features responsible for distinguishing between IBD, CD, or UC and between UC or CD.

This study provides a framework for combining the analysis of endoscopic and histological data with gene and genomic data to understand better the symptoms and cells associated with UC and CD. It was also shown that balancing data using the MSMOTE technique has an advantage over SMOTE and unbalance data in algorithm performance. While the best model may have been quite effective in classifying disease types with a high F1 and accuracy score, further concrete clinical studies and physician verification for each patient based on symptoms and history are recommended.

Acknowledgments. We would like to express our sincere appreciation and gratitude to everyone who contributed to this research Jacquelyn Cheun, PhD. – Capstone Professor, Nima Saeidi, Ph.D. Assistant Professor Department of Surgery Massachusetts General Hospital Harvard Medical School, Gautham Sridharan, Principal Scientist at Alnylam Pharmaceutical/ instructor at Tuft University.

References

Boyd, M., Thodberg, M., Viezic, M., Bornholdt, J., Vitting-Seerup , K., Chen, Y., . . . Sandelin, A. (2018, April 25). Characterization of the enhancer and promoter landscape of inflammatory bowel disease from human colon biopsies. *Nature communications, 9*(1), 1661. doi:10.1038/s41467-018-03766-z

Brownlee, J. (2016, April 15). *Boosting and AdaBoost for Machine Learning*. Retrieved June 12, 2021, from Machine Learning Mastery: https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/

Brownlee, J. (2016, April 15). *K-Nearest Neighbors for Machine Learning*. Retrieved June 12, 2021, from Machine Learning Mastery: https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/

Brownlee, J. (2016, April 15). *Logistic Regression for Machine Learning*. Retrieved June 12, 2021, from Machine Learning Mastery: https://machinelearningmastery.com/logistic-regression-for-machine-learning/

Brownlee, J. (2018, May 23). *Gentle Introduction to k-fold cross validation*. Retrieved June 12, 2021, from Machine Learning Mastery: https://machinelearningmastery.com/k-fold-cross-validation/

Brownlee, J. (2018, August 31). *How to Use ROC Curves and Precision-Recall Curves for Classification in Python*. Retrieved June 12, 2021, from Machine Learning Mastery: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

Cherney, K. (2020, September 30). *Stages of Crohn's Disease*. Retrieved June 12, 2021, from healthline: https://www.healthline.com/health/crohns-disease/stages

Cross, S. S. (1999). *Dataset of Observed Features on Endoscopic Colorectal Biopsies from Normal Subjects and Patients with Chronic Inflammatory Bowel Disease*

(Crohn's disease and Ulcerative Colitis). Department of Pathology, University of Sheffield Medical School.

Cross, S. S., & Harrison, R. F. (2002). Discriminant histological features in the diagnosis of chronic idiopathic inflammatory bowel disease: analysis of a large dataset by a novel data visualisation technique. *Journal of Clinical Pathology, 55*(1), 51-57. doi:10.1136/jcp.55.1.51

Cruz-Ramírez, N., Acosta-Mesa, H.-G., Barrientos-Martínez, R.-E., & Nava-Fernández, L.-A. (2006). Diagnosis of Chronic Idiopathic Inflammatory Bowel Disease Using Bayesian Networks. In J. Martínez-Trinidad, J. Carrasco Ochoa, & J. Kittler (Eds.), *Progress in Pattern Recognition, Image Analysis and Applications* (pp. 706-715). Berlin, Heidelberg: Springer Berlin Heidelberg.

Crypt distortion. (n.d.). Retrieved June 13, 2021, from My Pathology Report: https://www.mypathologyreport.ca/crypt-distortion/

Data and Statistics. (2020, August 11). Retrieved June 13, 2021, from Centers for Disease Control and Prevention: https://www.cdc.gov/ibd/data-statistics.htm

D'haeseleer, P. (2005, December). How does gene expression clustering work? *Nature Biotechnology, 23*(12), 1499-1501. doi:10.1038/nbt1205-1499

Dutta, P., Saha, S., Pai, S., & Kumar, A. (2020, January). A Protein Interaction Information-based Generative Model for Enhancing Gene Clustering. *Scientific Reports, 10*(1), 665. doi:10.1038/s41598-020-57437-5

Han, L., Maciejewski, M., Brockel, C., Gordon, W., Snapper, S. B., Korzenik, J. R., . . . Altman, R. B. (2008, March 15). A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. *Bioinformatics (Oxford, England), 34*(6), 985-993. doi:10.1093/bioinformatics/btx651

Hassan, A. M., & Omar, Y. M. (2020). Inflammatory Bowel Disease Classification Using Neural Network and Support Vector Machine. *International Research Journal of Advanced Engineering and Science, 5*(1), 216-221.

Hu, S., Liang, Y., Ma, L., & He, Y. (2009). MSMOTE: Improving Classification Performance When Training Data is Imbalanced. *2009 Second International Workshop on Computer Science and Engineering*, *2*, pp. 13-17. doi:10.1109/WCSE.2009.756

Inflammatory bowel disease (IBD). (n.d.). Retrieved June 13, 2021, from Mayo Clinic: https://www.mayoclinic.org/diseases-conditions/inflammatory-bowel-disease/symptoms-causes/syc-20353315

Jeggari, A., & Alexeyenko, A. (2017, March 23). NEArender: an R package for functional interpretation of 'omics' data via network enrichment analysis. *BMC Bioinformatics, 18*, 118. doi:10.1186/s12859-017-1534-y

Khorasani, H. M., Usefi, H., & Peña-Castillo , L. (2020, August 13). Detecting ulcerative colitis from colon samples using efficient feature selection and machine learning. *Scientific reports, 10*(1), 13744. doi:10.1038/s41598-020-70583-0

Li, B., & Lu, P. (2019, October 16). *SMOTE Azure Machine Learning*. Retrieved June 12, 2021, from Microsoft Docs: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/smote

Mahapatra, D., Schueffler, P., Tielbeek, J. A., Buhmann, J. M., & Vos, F. M. (2012). A Supervised Learning Based Approach to Detect Crohn's Disease in

Abdominal MR Volumes. In Y. Hiroyuki, D. Hawkes, & M. W. Vannier (Eds.), *Abdominal Imaging. Computational and Clinical Application* (pp. 97-106). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-33612-6_11

Manandhar, I., Alimadadi, A., Aryal, S., Munroe, P. B., Joe, B., & Cheng, X. (2021, January 13). Gut microbiome-based supervised machine learning for clinical diagnosis of inflammatory bowel diseases. *American Journal of Physiology-Gastrointestinal and Liver Physiology*. doi:10.1152/ajpgi.00360.2020

Mossotto, E., Ashton, J. J., Coelho, T., Beattie, R. M., MacArthur, B. D., & Ennis, S. (2017, May 25). Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. *Scientific reports*, 2427. doi:10.1038/s41598-017-02606-2

Ray, S. (2017, September 13). *Understanding Support Vector Machine(SVM) algorithm from examples (along with code)*. Retrieved June 27, 2021, from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

Smillie, C. S., Biton, M., Ordovas-Montanes, J., Sullivan, K. M., Burgin, G., Graham, D. B., . . . Regev, A. (2019, July 25). Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell, 178*(3), 714-730. doi:10.1016/j.cell.2019.06.029

Tan, T. Z., Ng, G. S., & Erdogan, S. S. (2006). A Neuropsychology-inspired Learning System for Human Uncertainty Monitoring. *9th International Conference on Control, Automation, Robotics and Vision*, (pp. 1-8). doi:10.1109/ICARCV.2006.345430

Troelsen, J., Petersen, A., Jensen, K., Gögenur, I., Thielsen, P., Seidelin, J., . . . Sandelin, A. (2018, April 25). Characterization of the enhancer and promoter landscape of inflammatory bowel disease from human colon biopsies. *Nature Communications, 9*(1), 1661. doi:10.1038/s41467-018-03766-z

What is IBD? (n.d.). Retrieved June 13, 2021, from UCLA Health: https://www.uclahealth.org/gastro/ibd/ulcerative-colitis-vs-crohns-disease#:~:text=The%20differences%20between%20ulcerative%20colitis,mixed%20in%20between%20inflamed%20areas

Widmann, M. (2020, August 4). *Cohen's Kappa: What It Is, When to Use It, and How to Avoid Its Pitfalls*. Retrieved June 12, 2021, from The New Stack: https://thenewstack.io/cohens-kappa-what-it-is-when-to-use-it-and-how-to-avoid-its-pitfalls/

Wood, T. (n.d.). *What is the F-score?* Retrieved June 12, 2021, from DeepAI: https://deepai.org/machine-learning-glossary-and-terms/f-score

Yoon, S., Kim, J., Kim, S.-K., Baik, B., Chi, S.-M., Kim, S.-Y., & Nam, D. (2019, May 9). GScluster: network-weighted gene-set clustering analysis. *BMC Genomics, 20*(1), 665. doi:10.1186/s12864-019-5738-6

Zhou, V. (2019, March 3). *Machine Learning for Beginners: An Introduction to Neural Networks*. Retrieved June 27, 2021, from Victor Zhou Blog: https://victorzhou.com/blog/intro-to-neural-networks/