Spring 2020

Inflammatory Bowel Disease Diagnosis Using Metagenomic Classification

Michael Riggle
Western Kentucky University, michaelpriggle@gmail.com

INFLAMMATORY BOWEL DISEASE DIAGNOSIS
USING METAGENOMIC CLASSIFICATION

A Thesis
Presented to
The Faculty of the School of Engineering and Applied Sciences
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science

By
Michael Riggle

May 2020

INFLAMMATORY BOWEL DISEASE DIAGNOSIS USING METAGENOMIC CLASSIFICATION

Date Recommended 6/30/20

Dr. Qi Li, Director of Thesis

Xing, Guangming Digitally signed by Xing, Guangming
Date: 2020.07.01 15:08:37 -05'00'

Dr. Guangming Xing

Zhonghang Xia Digitally signed by Zhonghang Xia
Date: 2020.07.01 14:45:58 -05'00'

Dr. Zhonghang Xia

Ranjit T. Koodali Digitally signed by Ranjit T. Koodali
Date: 2020.07.14 08:42:00 -05'00' 7/14/20

Dean, Graduate School Date

Dedicated to Mom, Dad, Matthew, and Jennifer

Thank you for all the love and support.

ACKNOWLEDGMENTS

The completion of this thesis would not have been possible without the support of many individuals. First, I would like to thank my thesis advisor, Dr. Qi Li, for his guidance, knowledge, and advice not only throughout the thesis writing process but also during the many classes I had him as a professor.

Next, I would like to thank Dr. Guangming Xing and Dr. Zhonghang Xia for serving on my committee, for providing me guidance with my thesis, and the support they have provided both inside and outside the classroom during my time at WKU.

I would like to thank the rest of the computer science faculty at WKU that taught me in both my undergraduate and graduate careers. I enjoyed my time in the computer science program and I am thankful for the opportunities and challenges that shaped me into a better computer scientist.

I would also like to thank my high school programming teacher, Ms. Connie Sieja, for helping to cultivate a love for coding and giving me the tools to further the pursuit of computer science.

I would like to express my gratitude to the Inflammatory Bowel Disease Multi'omics Database and the Integrative Human Microbiome Project for all the research they have done with IBD and provide their findings and datasets publicly for other researchers like myself to use.

Lastly, I would like to thank my family and friends for the support throughout my academic career, especially through the thesis writing process.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

INFLAMMATORY BOWEL DISEASE DIAGNOSIS USING METAGENOMIC CLASSIFICATION

Michael Riggle May 2020 59 Pages

Directed by: Qi Li, Guangming Xing, and Zhonghang Xi

School of Engineering and Applied Sciences Western Kentucky University

Inflammatory bowel disease (IBD) is a set of disorders that involve chronic inflammation of digestive tracts, e.g., Crohn's disease (CD) and ulcerative colitis (UC). Millions of people around the world have inflammatory bowel disease. However, it is still difficult to treat IBD due to its unknown cause. In fact, accurately diagnosing inflammatory bowel disease (IBD) can be very challenging too since some of IBD symptoms can mimic those of other conditions. In this work, we apply classification methods to help improve the success rate of diagnosis. We study four formulations of IBD classification: i) IBD and non-IBD (binary classification), ii) CD and non-IBD (binary classification), iii) UC and non-IBD (binary classification), and iv) UC, and non-IBD (ternary classification). We have applied a number of classification methods, including decision tree, Naive Bayes, K-nearest neighbor, and rule-based classifier, to the two IBD classification problems using a metagenomic dataset collected from stool samples. Our study shows that a rule-based classifier achieves the best combination of classification accuracy and readability. We also explored the roles of attributes in the diagnosis of IBD based on interpretation of learned models. Studying the importance of specific attributes could lead to a better understanding of IBD by either discovering new connections or reinforcing known ones.

Chapter 1. Introduction

Inflammatory bowel disease (IBD) is a group of disorders that cause chronic inflammation of the gastrointestinal (GI) tract that causes an array of symptoms. Over 3.5 million people are diagnosed with IBD and that number is still increasing, not only in North America and Europe where is it most common, but also around the world [1]. There are two main types of IBD with the majority of diagnoses: Crohn’s disease (CD) and ulcerative colitis (UC). While CD and UC are very similar, they do have specific difference. CD can affect anywhere along whole GI tract (from mouth to anus) with damaging inflammation appearing in patches and with the possibility of reaching multiple layers of the GI tract wall, while UC can only affect the large intestine (colon) and sometimes rectum, with damage continuous and inflammation is only in the innermost layer of the colon lining [2].

CD and UC are very similar with their symptoms and most (but not all) treatment options. However, if surgery is required, the type of surgery and area operated on are different due to the difference in inflammation. Due to the continuous damage of UC, it is far easier to perform a removal surgery to remove only the parts of the colon that are affected. CD’s damage often cannot be simply removed because it only appears in scattered patches and/or it may appear in areas that cannot be removed. [3]. For this reason, the distinction between CD and UC in IBD diagnosis is important so that medical professionals and patients are aware of all options open to them.

While IBD is not fatal (unless there are major complications), it has a large impact on the lives of those who have the disease, especially if remission cannot be reached. Common symptoms include but not limited to abdominal cramps and pain, fatigue, diarrhea, and bloody stools. IBD also increases chances of having many other diseases or disorders such as, cardiovascular disease, respiratory disease, cancer, arthritis, weak or failing kidneys, any liver condition, and ulcer [4]. IBD can bring additional challenges from the symptoms and risk factors. With the lack of public understanding of IBD, patients can be treated with insensitivity to their condition, either by not understanding the severity of their condition or not believing their condition at all. This can cause them to keep their condition to themselves and fail to reach out for support which in turn leads to stress, anxiety, depression, and/or other mental health problems. The stress and cost of treatment and eve bathroom access can also provide more challenges to patients [5].

The previously mentioned symptoms and possible other condition that IBD patients can have may lead to a misunderstanding that a patient has other disorders such as Irritable Bowel Syndrome, celiac disease, tuberculous enteritis, duodenal ulcer, appendicitis, anal fistula, enterocolitis, hemorrhoids, rectal varix, and rheumatoid arthritis [2] [5]. In summary, IBD can be difficult to diagnose because its cause is unknown, and it can be easily diagnosed as other diseases with similar appearing conditions. This can cause many patients to go untreated while for medical professionals determine what might be causing their symptoms and provide a correct diagnosis. A study performed by Yong Hoon Kwon and Yong Joo Kim show that the average diagnostic time lag in children with CD was 3.36 months, and with UC was 2.2 months but this time can increase by month or years with an incorrect diagnosis [5]. This can cause patients to suffer with symptoms longer than necessarily and receive treatment for a condition that they do not have which may harm them more. There are a few tests that medical professionals order which all have varying goals, success, and burden on the patients

Although the exact cause of IBD is unknown, there are some known factors that increase the chance of someone having it and changes in gut microbiota are seen as well. These factors include, age, race/ethnicity, family history, smoking, and environment [6] [7]. The change in the gut microbiota is highly debated because it is unknown whether the change is the cause of the IBD or the consequence of it [8]. This debate between change and consequence of the gut microbiota change causes many different paths in treatments such as chemotherapy, complementary and alternative medicine, traditional Chinese herbal medicine, prebiotics, probiotics, synbiotics, and fecal microbiota transplantation. Some of these tests include, general blood test, stool test, medical imaging, colonoscopy, and colon biopsy [9]. Despite the variety of test, the colonoscopy is considered to the benchmark for monitor any IBD activity because of its accuracy and the amount of the information that gained in comparison to the other tests. The colonoscopy has an 89% accuracy of diagnosis of CD or UC [10]. With the proper medication, many patients find remission and experience little to no symptoms.

With the complexity of medical diagnosis and the human body, it can be quite difficult for humans to find the connects between the conditions of the human body and varies disorders. All the different conditions in the human body from DNA to heart rate to abundance of bacteria in a patient stool can be measured and recorded to produce large dataset that focus on the wellbeing of the body. With the use of machine learning techniques, these datasets can be processed to reveal knowledge the body that were not previously known and possibly quite hard for humans to detect on their own. While machine learning technique cannot replace the knowledge, intuition, and experience of medical professionals, they can be used as another tool to extend our understanding of the problem. Just like any other tool, there is no tool that works for all problems. This is why various machine learning techniques should be tested to find which one is best suited for a specific problem or set of problems.

This thesis discusses the diagnosis of IBD using classification methods while looking at three binary classifications (IBD and nonIBD, CD and nonIBD, UC and nonIBD) and ternary classification (CD, UC, and nonIBD). The models used include C4.5 Decision Tree, Naïve Bayes, k-Nearest Neighbor, RIPPER, and Decision Table. The ensemble methods of bagging and boosting are also used on the C4.5 decision tree and the RIPPER algorithm. The use of these classification models to diagnose IBD using a Metagenomic dataset collected from stool samples. Waikato Environment for Knowledge Analysis (WEKA) is used in the implementation of the classification models. The findings of this thesis show that the RIPPER algorithm has great potential for diagnosing IBD because of its interpretability, a higher classification accuracy than a colonoscopy, and the rule set lines up with some known links between IBD and certain bacteria.

Chapter 2. Literature Review

The use of data mining models and machine learning techniques in the medical field is not a new area of research and there is research that looks directly at diagnosing IBD. This allows for the comparisons of results, classification models, and datasets and also a comparison to the current test standards. While the combination of the classification models and metagenomic data is what being studied in the theis, the study of metagenomic data from stool samples has been used to learn more about a patient's microbiota and test for infection.

2.1 Current IBD Testing Methods

Isabelle Noiseux et al. discusses the various test methods being used today and studied the cost, refusal rates, comfort and fear levels, knowledge, etc. that patient associate with the different tests in their study, "Inflammatory bowel disease patient perceptions of diagnostic and monitoring tests and procedures" [9]. They collected survey data through the Crohn's and Colitis Canada association about the five common tests used to test for IBD: general blood test, stool test, colonoscopy, colon biopsy and medical imaging. The survey received 210 responds where 145 had CD and the other 65 had UC. The study first looked at what tests were requested the most and were refused the most with both being the general blood test even when it provides the least information. They continued by investigating why patients refuse specific tests. The study continues by looking at patients' comfort, understanding of the tests, and what information about the test was provided by the medical professional. This information can be used to help decide what tests to order, where medical professional can improve when interacting with patients, and understand how patients view the tests. Patients appeared to be most comfortable with the stool sample and medical professionals need to provide more information about risks of having false positive or negative tests.

2.2 Data Mining and Medicine

Elahe Parva et al. discusses the emerges of data mining in the medical field, especially in emergency medicine in their paper, “The Necessity of Data Mining in Clinical Emergency Medicine; A Narrative Review of the Current Literature” [11]. Parva discusses the complexity of the body with its many systems and subsystems is much like the complexity and connectivity of economic and military problems that already use machine learning and statistics to effectively solve complex problems. Data mining would not only help with understanding the body and diagnosis problems but also reduce the waste of medical resources. Parva identifies the following areas of medical science as potential places data mining analysis can be applied [11]:

- Identifying the complex mechanisms of different body subsystems and their interactions with each other
- Identifying people who are at risk for diseases of a genetic predisposition or caused by environmental factors
- Identifying disease mechanisms and their interactions with the problems of the body
- Determining disease prognoses, and facilities management
- Establishing decision support systems to make the best decision, especially when the disease is multi-factorial, when more factors are

involved in determining the course of the disease, in emergencies, or in acute phases of a disease

- Evaluating diagnostic and treatment tasks and relationships and identifying shortcomings and capabilities
- Finding the best screening methods for diseases and injuries, particularly for patients in critical conditions.

The paper continues by discussing the studies that have already applied data mining in various medical areas including, cancer, prognosis of patients after lung transplantation, determining diseases and facilities management, and decision support system.

2.3 Endoscopy and Histology

E. Mossotto et al. studies the use of ensemble learners, linear discriminant analysis and support vector machine (SVM) for IBD diagnosis in their paper "Classification of Paediatric Inflammatory Bowel Disease using Machine Learning" [12]. Mossotto's research uses data obtained from endoscopy and histology (study of the microscopic structure of tissues) at initial diagnosis from 287 IBD patients (178 CD, 80 UC, and 29 unclassified IBD (IBDU)). This study has no healthy controls to provide a baseline for people with IBD, which means that this study only provide insight in the difference between types of IBD. This study investigates both unsupervised and supervised learning. The unsupervised clustering did not should significant separation between types of IBD. Hierarchical clustering is able to group most patients into 4 groups but there is not significant different between groups. The following models where tested in order to decision which model to optimize; simple tree, medium tree, complex tree, linear discriminant analysis, linear SVM, quadratic SVM, cubic SVM, boosted trees, and bagged trees. Out of these models, linear discriminant analysis had the best accuracy with 81.0% followed closely by linear SVM which was a half percent lower. The SVM was used for optimization for its adaptability and interpretability. The optimized SVM when looking at both the endoscopy and histology data had the best accuracy with 82.7%.

2.4 Promteomic Mass Spectra

Pierre Geurts al et. perform a study in "Proteomic mass spectra classification using decision tree-based ensemble methods" [13]. This study's goal is to propose a flexible method to analysis and learn from proteomic mass spectra with the hopes of a framework that would be able to diagnosis multiple diseases. This study uses rheumatoid arthritis and IBD as proof of concept with the hopes of expanding to other diseases. Using rheumatoid arthritis and IBD have some shared characteristics and can treated with the same medication [14]. This demonstrates that similar problem, but different problems can be tackled by in a similar way. Geurts used the following models: single decision tree, bagging, random forests, extra-trees, boosting, k-nearest neighbors, and support vector machine. This paper uses a mass spectrometry data that generates protein profiles from body fluids with 240 instances with both IBD patients and healthy controls. Their experiments found that the model with the best error rate for the diagnosis of IBD was achieved by extra-tree with a 10.02% error rate with the best discretization. This study also investigates attribute importance ranking and biomarker selection. The attribute selection by boosting (r = 1%) decreasing the error rate to 6.68%.

2.5 Metagenomic Stool Sample

Federica Gigliucci al et. looks at metagenomic data from stool samples to identify patients that are positive for Shiga toxin-producing *Escherichia coli* (STEC) in "Metagenomic Characterization of the Human Intestinal Microbiota in Fecal Samples from STEC-Infected Patients" [15]. The study shows that the subjects' microbiota underwent a significant change between subjects with and without the STEC infection and even identify specific bacteria that are linked to the infection. This study has samples from patients with diarrhea, patients after 2 weeks from the restoration of the normal intestinal function, healthy subject, and patients with CD. The promising results of this study set a precedence for analysis the metagenomic data of stool sample when diagnosing diseases and infections that have major effect on the microbiota of a patient, including CD.

Chapter 3. Methodology and Dataset

These experiments use the Metagenomic dataset was collected and distributed by the Inflammatory Bowel Disease Multi'omics Database (IBDMDB) [16] as a part of the Integrative Human Microbiome Project (HMP2) [17]. Metagenome is the complete sequencing of all genetic material in an environment. In this dataset, the environments are the stool samples of each participant. The analysis of the metagenomic data of a stool sample can give an insight into the microbiota of the participant without the having an endoscopy, like a colonoscopy, performed. The study has 132 participants that are collected from three pediatric facilities (Cincinnati Children's Hospital, Massachusetts General Hospital (MGH) Pediatrics, and Emory University Hospital) and two adult facilities (MGH and Cedars-Sinai Medical Center). Multiple samples where taken from the participants creating a dataset has 1338 instances with 599 CD diagnoses, 375 UC diagnoses, and 364 nonIBD diagnoses as the control group of the study. The dataset has 1453 attributes with one class attribute after preprocessing. The dataset comes in two part: metadata and metagenomic data.

Table 1: HMP2: Metagenomes Dataset

Dataset Name	**Number of Instances**	**Number of Attributes**	**Number of Classes**	**Class Distribution**
HMP2: Metagenomes	1338	1454	3	CD – 599 UC – 375 nonIBD – 364

The metagenomic data in the dataset is sequences of bacteria, viruses, and other microbes found in a participants' stool sample while ignoring any of the human RNA for the participant. This RNA sequence data is ran through on a Anadama pipeline to create files that records the concentrations of the bacteria, viruses, and other microbes both as specific species and larger classifications [16]. These concentrations values are broken by according to the taxonomy of the microbes by not only looking at the concentration of specific species but also kingdom, phylum, class, order, family, genus, and trinomen (subspecies). There were about 540 different species identified but with filtering out of any species with less than 0.01% abundance, they narrowed the study do to 109 species [16].

The metadata contents attributes that were collected by having each participant fill out a form asking questions about their health, recent diet, and general information about themselves (age, sex, race, etc.). Some of these attributes were removed from the dataset during preprocessing because they had the same value for every data point and/or they were included for clerical purposes only and do not have any effect on the diagnosis (ex: Name of the testing center). The attribute "Age at diagnosis" was also remove because only the participants that have IBD have a diagnosis age. To allow for the both the binary and ternary classifications, the diagnosis attributes must make a new dataset where the diagnosis attributes was processed where any CD and UC labels to IBD, when looking at IBD and non IBD. Two other datasets were created by removing rows for either CD or UC, so comparisons between CD and nonIBD and between UC and nonIBD.

The models that this paper focuses on is C4.5 decision tree, RIPPER, Decision Table, Naïve Bayes, and k-nearest neighbor built-in models in WEKA. The ensemble learning methods Bagging and Boosting are applied on C4.5 decision tree and RIPPER.

Each model is ran with the default parameters in WEKA including using 10-fold cross-validation for their validation unless otherwise specified.

The C4.5 decision tree (or J48 in WEKA) creates a binary tree that can also ignore missing values [18]. It selected for this study because of its use in the simple studies, its simplicity, and readability. The RIPPER algorithm (or JRip in WEKA) and Decision Table were selected for their readability and to determine the effectiveness of rule-based classifiers on this diagnosis problem. The JRip classifier or sometimes IREP using reduced error pruning along with divide and conquer method of creating greedy rule set one rule at a time [19]. The Decision table or sometimes Decision table majority (DTM) is created by finding the optimal set of features that provide the highest accuracy [20]. Naïve Bayes was selected because it is known for being applied to medical diagnosis problems [21]. Naïve Bayes classifier provide a simple approach to calculating probabilities by counting the frequency and combinations of values with the assumption that the attributes are independent of the given class [18]. K-nearest neighbor (or IBk in WEKA) was selected for its simplicity allows for easy implementation and explanation to those who have not used the model before. When classifying new data points, k-nearest neighbor determines what the closer existing data points (or neighbors) are and determine the new point class based on the classes of its neighbors [22].

The Bagging and Boosting ensemble learning models were picked because they are two of the most common ensemble methods used and to draw comparison with the related studies which also used bagging and boosting. Bagging (also known as bootstrap aggregating) generates multiple versions of the selected method using bootstrap replicates to make an aggregated predictor [23]. Boosting (or AdaBoost in WEKA) is used to “boost” the accuracy of other algorithm by running the classifier multiple times with various distributions over the training data and then combines the runs to make a single composite classifier [24].

The decision tree and RIPPER are especially focused on due to their better readability which allows for additional information beyond their statistics (accuracy, ROC, etc.). The models can also be evaluated by investing if the models created make logical sense with the attributes used. Also, it allows users that might not be as familiar with data models, such as medical professions, to understand the model while using it for diagnosis. Other models such as neural networks were considered for this paper but since the dataset has discrete and non-discrete attributes, models like neural networks are not suited for this type of dataset.

All of the models were tested on five subsets of the attributes: all, metagenomic, meta, race, and sex. These subsets help to determine the importance of varies attributes in the different classification problems and in the varies models used. The attribute race was selected because it is known risk factor. However, sex was selected because females are more likely to have CD even if the sex patient is not a specific risk factor [4]. The subset *all* consists of all the attribute not remove with preprocessing, *metagenomic* consists only of the numeric concentrations values from the metagenomic data and the diagnosis, *metadata* consists of all attributes from the metadata, *race* consists of the metagenomic values, the diagnosis, and race, and *sex* consists of the metagenomic values, the diagnosis, and sex. The *metadata* subset is unlikely to provide a good diagnostic tool, but it may lead to a better understanding of what attributes from the metadata are important in the diagnosis problem.

Chapter 4. Binary Classification: IBD and nonIBD

This section discusses the binary classification of instances as either IBD or nonIBD. This binary classification focus on whether a patient has IBD, in general, so future tests would have to be performed to given them a more specific diagnosis. This section will compare accuracies against those obtained by Geurt's study [13]. A comparison is not been made to Mossotto's study because the diagnosis is only between CD and UC without healthy controls [12].

4.1 Initial Results

Figure 1 shows a comparison between the different models on the different subsets on the IBD/nonIBD dataset. The dataset consists of 974 instances of IBD and 364 instances of nonIBD.

4.1.1 Decision Table

At first glance, it appears that the decision table using the *all* subset is the best model option for the IBD and nonIBD model. The decision table has good readability and has the best accuracy of the models and subsets, but the rules created show that it is not a good option for this dataset. The table created with the *all* use following attributes: "consent_age", "Immunosuppressants (e.g. oral corticosteroids)", "In the past 2 weeks have you used an oral contrast", "General wellbeing", "Bowel frequency during the day", and "*Haemophilus pittmaniae*". Generally a smaller decision table usually means a better decision table, but asking for a patients age, medications they are using, their general wellbeing, and looking at concentration of one bacteria, the is not going to be enough information to diagnosis someone with a complex disease like IBD. The decision table using the *metadata* subset also has quite high accuracy, but it uses the same attributes as the *all* subset minus the "*Haemophilus pittmaniae*" attribute. The decision table has more reasonable attribute selection for the other subsets, however, these tables are not as attractive as a choice because they have lower accuracy than the models with similar accuracy.

4.1.2 K-Nearest Neighbor

After the decision table, k-nearest neighbor has high accuracy in comparison to other models. The k-nearest neighbor with all the different subsets of this dataset also outperformed the k-nearest neighbor from Geurts's study which has an error rate of 20.21% (or accuracy of 79.79%) [13]. While the accuracy of the model is the highest, the model has low readability. While the concept of nearest neighbor is simple and easy for people to understand, it is hard to actually gain additional information from the model. This means that no addition information can be gathered from the model which in turn means fewer medical professionals would trust the model's diagnosis.

4.1.3 Naïve Bayes

The Naïve Bayes model does not perform very well for all the different subsets with the exception of the *metadata* subset. This model appears to not be equipped to work with the metagenomic abundance data because the accuracy of this model increases with the two subsets that include the metadata (*all* and *metadata* subsets). However, the specific metadata attributes of sex and race appear to have no effect on the model since the accuracy of the *metagenomic*, *sex*, and *race* subsets are the same. Despite the Naïve

Bayes model being popular for medical diagnosis problems, it is not a good fit for this specific problem.

4.1.4 C4.5 Decision Tree

The C4.5 decision tree has the best accuracy when classifying IBD and nonIBD when looking at the *sex* subset. After the decision table and k-nearest neighbor models, C4.5 decision tree has the highest accuracy but only when using the *sex* subset. The C4.5 decision tree is quite consistent across the different subsets with the exception of the *metadata* subset, since the accuracy between subset has less than a percent difference. The good readability of the decision tree also makes this model more trusted because the user knows why the model has made its prediction. The C4.5 decision using the *sex* outperforms the single tree in Geurts's study, with a 26.67% error rate (or accuracy of 73.3%) [13], and it also outperforms a colonoscopy by 3.75% [10]. However, the accuracy of the best method in Geurts' study which is using attribute selection by boosting (r = 1%) outperformed the C4.5 decision tree but only by about 0.57% [13].

4.1.5 RIPPER

The RIPPER algorithm also has merit for the classification between IBD and nonIBD. When using the *all* subset, the accuracy is the highest after the decision table and k-nearest neighbor. Since the RIPPER is a rule-based classifier, the readability is excellent even for individual with little to no experience with the model itself. The size of the RIPPER algorithm increases its attractiveness in comparison with the C4.5 decision tree. The RIPPER algorithm has 11 rules while the C4.5 decision tree has a size of 89 with 45 leaves. The receiver operating characteristic area (ROC area) of the RIPPER algorithm (90.4%) is higher than the ROC area of the C4.5 decision tree (89.4%).

Because of these factors, RIPPER algorithm is a better model for the IBD and nonIBD classification problem, especially with it still outperforming the colonoscopy by 3.45% [10].

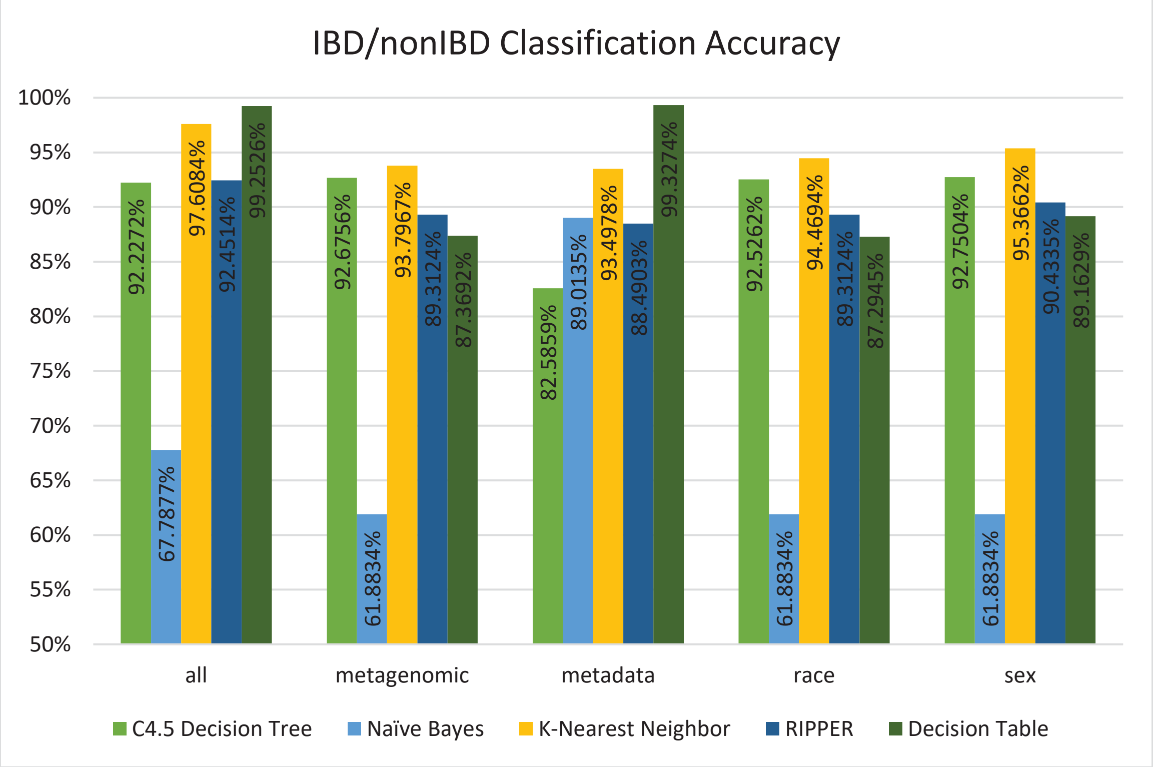

Figure 1: IBD/nonIBD Classification Accuracy Chart

4.2 Ensemble Learning

The ensemble methods improved the accuracy of both the C4.5 decision tree and the RIPPER algorithm. The increase between the single models and the bagging methods has a similar increase for both models with the RIPPER algorithm having a larger increase. The decision tree has an increase of about 3.66% while the RIPPER algorithm of about 4.26% while using the bagging method. The boosting, however, has a larger impact on the RIPPER algorithm's accuracy then the accuracy for the decision tree (with about 6.05% and 4.71% change respectively). This study's bagging and boosting for trees outperform Geurt's study who's bagging tree error rate is 16.25% (or accuracy of 83.75%) and boosting tree's error rate is 11.46% (or accuracy of 88.54%) [13].

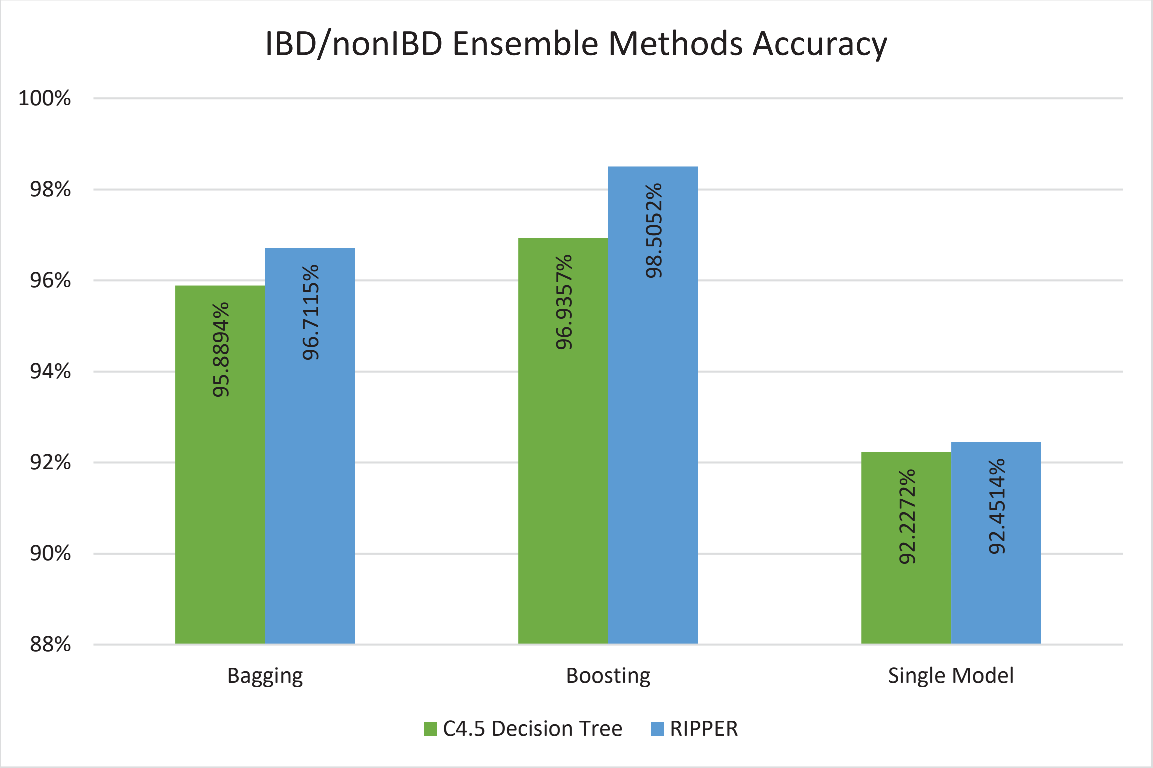

Figure 2: IBD/nonIBD Ensemble Methods Accuracy Chart

4.3 Accuracy Improvement

In an attempt to improve the accuracy of the RIPPER algortihm, the number of optimzation runs performed during the optimization stage for both the single model and for ensemble learning. The default number of optimzation runs in WEKA is 2.

4.3.1 Best Model - RIPPER

We see little change when only a few extra runs are added to the optimization stage. When changed from 2 to 3, 4, or 5, the accuracy, false positive rate, and ROC Area only changes by less then a percent and with the addition of one rule. However, if the number of optimzation runs is increased to 10 or 20, the accuracy and ROC Area is inceased significant and false positive rate decreases significantly. Using 10 runs has a sightly smaller false positive rate then using 20 runs but using 20 runs has better accuracy then using 10 runs (by 0.23%) and has two less rules which decrease the complex of the rule set. For different optimization runs performed, 20 runs is the best choice for binary classification. Using 20 optimization runs with the RIPPER algorithm outperforms the accuracy of a colonoscopy by almost 5% [10].

Table 2: RIPPER Optimization for IBD/nonIBD Classification

Number of Optimizations Runs	**Accuracy**	**FP Rate**	**ROC Area**	**Number of Rules**
2	92.45%	12.1%	90.5%	11
3	92.45%	11.9%	90.5%	12
4	91.93%	12.3%	90.2%	12
5	92.90%	11.4%	90.8%	12
10	93.72%	8.4%	93.0%	12
20	93.95%	8.5%	93.0%	10

4.3.2 Improved Ensemble Learning

To continue improving accuracy, the RIPPER algortihm with 20 optizimation runs is used with the ensemble learning methods of bagging and boosting. As show above, bagging improves the 2 run RIPPER algortihm by about 4.26% and boosting improves it by 6.05%. From the 2 run RIPPER alogrithm, there is an increase of about 3.36% when using bagging and an increase of about 5.38% when using boosting on the

20 run RIPPER algorithm. While the ensemble learning with the increased number of optizimation runs has a high accuracy, it has a smaller increase of the accuracy.

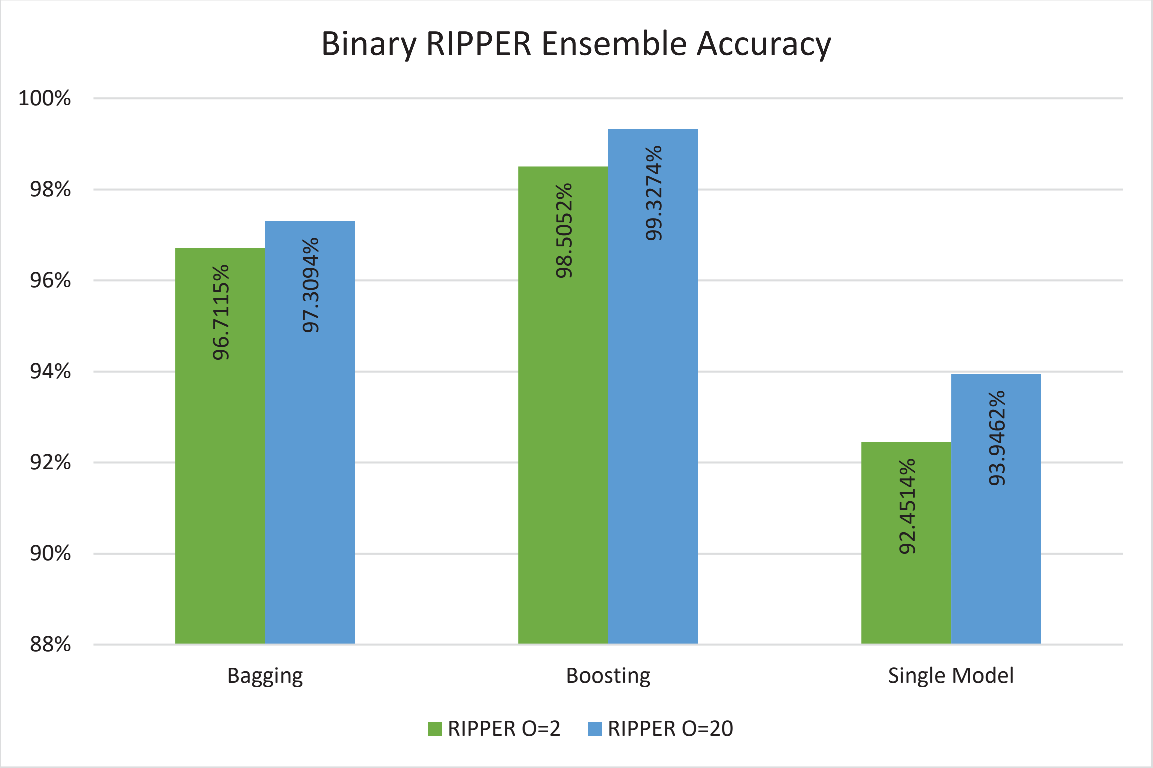

Figure 3: IBD/nonIBD Improved RIPPER Ensemble Accuracy

Chapter 5. Binary Classification: CD and nonIBD

This section discusses the binary classification of instances as either CD or nonIBD. This binary classification focus on whether a patient has CD, so future tests would have to be performed to see if a patient has UC if the test comes up negative. This section will compare accuracies against those obtained by Geurt's study but keep in mind that their study is comparing IBD and nonIBD [13]. A comparison is not been made to Mossotto's study because the diagnosis is only between CD and UC without healthy controls [12].

5.1 Initial Results

Figure 4 shows a comparison between the different models on the different subsets on the CD/nonIBD dataset. The dataset consists of 599 instances of CD and 364 instances of nonIBD.

5.1.1 Decision Table

At first glance, it appears that the decision table using the *all* subset is the best model option for the CD and nonIBD model. The decision table has good readability and has the best accuracy of the models and subsets, but the rules created show that it is not a good option for this dataset. The table created with the *all* use following attributes: "consent_age", "Have you ever had bowel surgery?" "General wellbeing", and "*Bacteroides dorei*". Generally a smaller decision table usually means a better decision table, but asking for a patients age, whether they have had bowel surgery, their general wellbeing, and looking at concentration of one bacteria, the is not going to be enough information to diagnosis someone with a complex disease like IBD. The decision table using the *metadata* subset also has quite high accuracy, but it uses the same attributes as the *all* subset minus the " *Bacteroides dorei*" attribute. The decision table has more reasonable attribute selection for the other subsets, however, these tables are not as attractive as a choice because they have lower accuracy than the models with similar accuracy.

5.1.2 K-Nearest Neighbor

After the decision table, k-nearest neighbor has high accuracy in comparison to other models. The k-nearest neighbor with all the different subsets of this dataset also outperformed the k-nearest neighbor from Geurts's study which has an error rate of 20.21% (or accuracy of 79.79%) [13]. While the accuracy of the model is the highest, the model has low readability. Due to the difficulty of reading the k-nearest neighbor, no addition information can be gathered from the model which in turn means fewer medical professionals would trust the model's diagnosis.

5.1.3 Naïve Bayes

The Naïve Bayes model does not perform very well for all the different subsets with the exception of the *metadata* subset. This model appears to not be equipped to work with the metagenomic abundance data because the accuracy of this model increases with the two subsets that include the metadata (*all* and *metadata* subsets). However, the specific metadata attributes of sex and race appear to have no effect on the model since the accuracy of the *metagenomic*, *sex*, and *race* subsets are the same. Despite the Naïve Bayes model being popular for medical diagnosis problems, it is not a good fit for this specific problem.

5.1.4 C4.5 Decision Tree

The C4.5 decision tree has the best accuracy when classifying CD and nonIBD when looking at the *all* subset. After the decision table and k-nearest neighbor models, C4.5 decision tree has the highest accuracy but only when using the *all* subset.

The C4.5 decision tree is quite consistent across the different subsets with the exception of the *metadata* subset, since the accuracy between subset has less than a percent difference. The good readability of the decision tree also makes this model more trusted because the user knows why the model has made its prediction. The C4.5 decision tree using the *all* outperforms the single tree in Geurts's study by about 17.77% [13], and it also outperforms a colonoscopy by 2.07% [10]. However, the accuracy of the best method in Geurts' study which is using attribute selection by boosting (r = 1%) outperformed the C4.5 decision tree but only by about 1.41% [13].

5.1.5 RIPPER

The RIPPER algorithm also has merit with the classification between IBD and nonIBD. When using the *all* subset, the accuracy is the highest after the decision table and k-nearest neighbor. Since the RIPPER is a rule-based classifier, the readability is excellent even for individual with little to no experience with the model itself. The size of the RIPPER algorithm increases its attractiveness in comparison with the C4.5 decision tree. The RIPPER algorithm has 11 rules while the C4.5 decision tree has a size of 85 with 48 leaves. Because of these factors, RIPPER algorithm is a better model for the CD and nonIBD classification problem, especially with it still outperforming the colonoscopy by 1.55% [10].

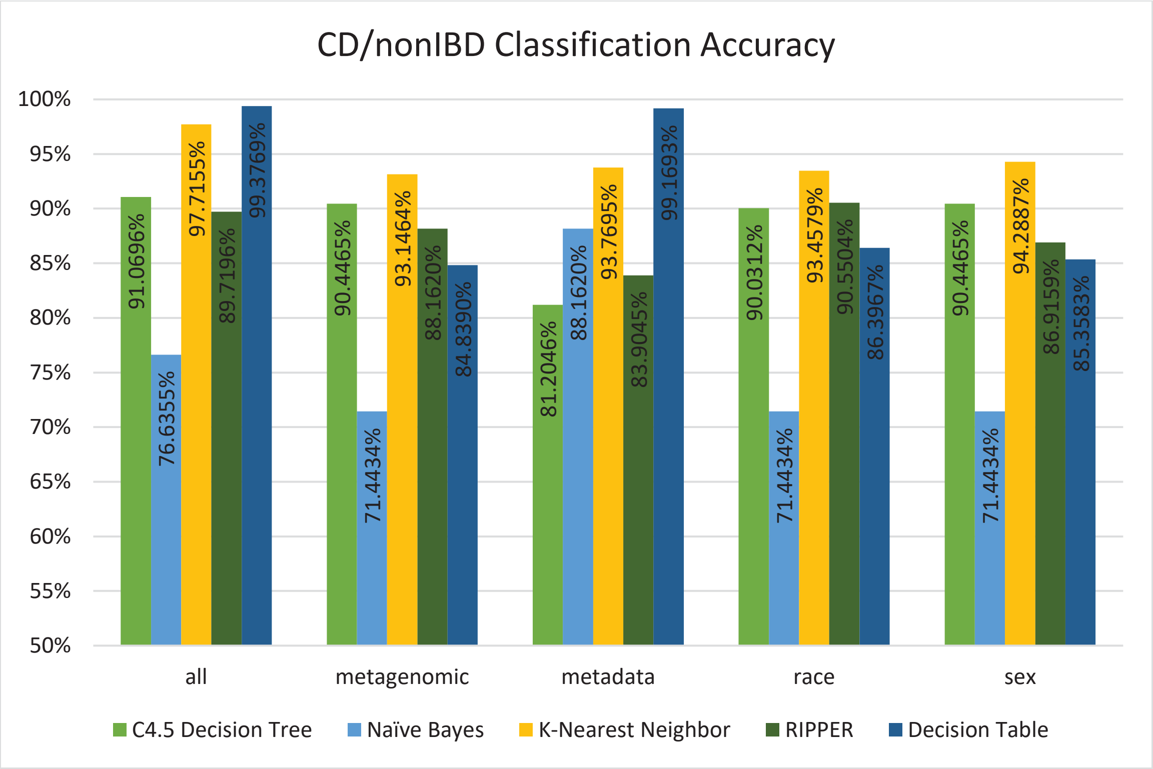

Figure 4: CD/nonIBD Classification Accuracy

5.2 Ensemble Learning

The ensemble methods improved the accuracy of both the C4.5 decision tree and the RIPPER algorithm. The boosting method performs better than bagging method. The C4.5 decision tree has a larger accuracy than RIPPER when using the boosting method, but the RIPPER algorithm performs better than C4.5 decision tree in the bagging method. In the bagging method, the decision tree has an increase of about 3.32% while the RIPPER algorithm of about 5.82%. The boosting, however, has a larger impact on the RIPPER algorithm's accuracy then the accuracy for the decision tree (with about an 7.17% and 6.65% respectively), but the C4.5 decision tree has larger accuracy then the RIPPER algorithm. This study's bagging and boosting for trees outperform Geurt's study by about 10.64% and outperforms the boosting by about 9.18% [13].

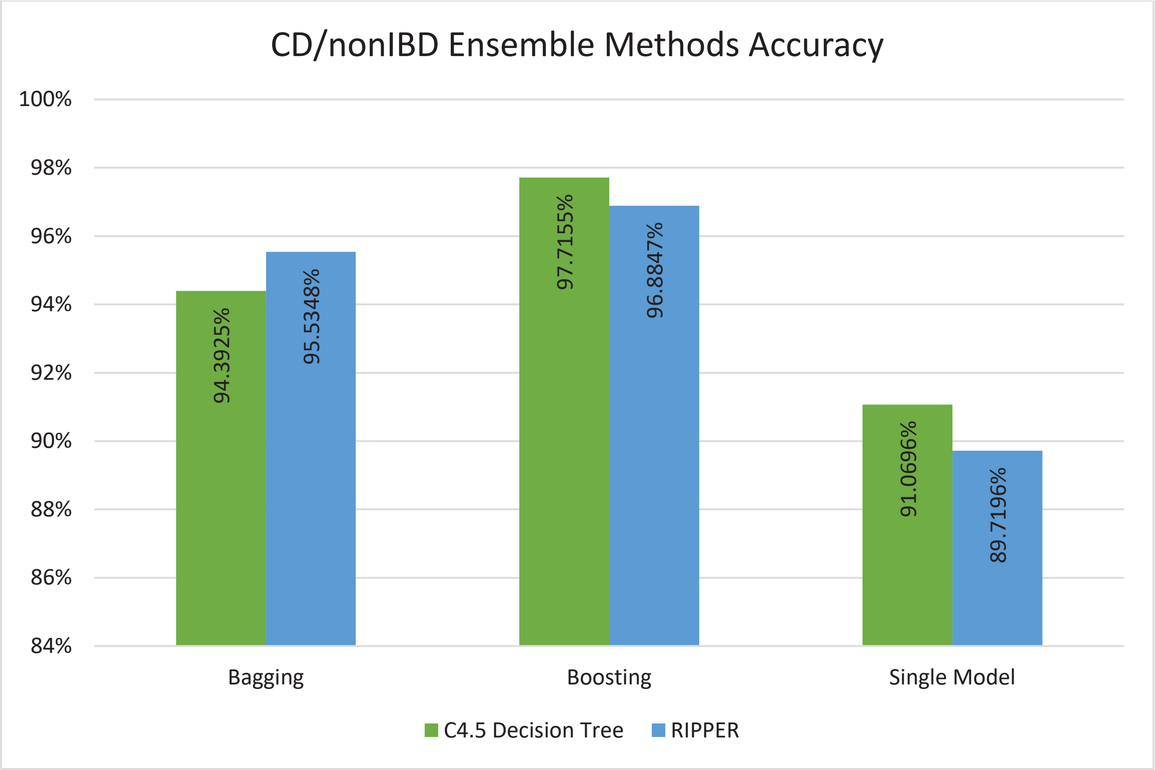

Figure 5: CD/nonIBD Ensemble Methods Accuracy

5.3 Accuracy Improvement

In an attempt to increase the accuracy of the C4.5 decision tree, the confidence factor related to pruning the tree has been changed. The default value in WEKA for the confidence factor is 0.25. The decrease of this number increases the amount of pruning. In this experiment, the confidence factor will only be lowered to guard against overfitting.

5.3.1 Best Model – C4.5 Decision Tree

Very little change occurs in the accuracy when the confidence factor is decreased. The accuracy does increase between the confidence factor 0.25 and 0.20 or 0.25and 0.15. The decision trees with confidence factors 0.20 and 0.15 have the same accuracy with 91.1734%. The confidence factor of 0.15 provides a slightly better tree with the ROC

Area increasing by 0.1%, the number of leaves decreasing by 7, and the size decreases by 9. The confidence factors of 0.10 and 0.05 both have a decrease in accuracy with 0.05 having the worse accuracy of the two trees created. The best accuracy made by the

Table 3: C4.5 Decision Tree Optimization for CD/nonIBD Classification

Confidence Factor	Accuracy	FP Rate	ROC Area	Number of Leaves	Size
0.25	91.07 %	10.2%	93.1%	48	85
0.20	91.17%	10.0%	93.0%	48	85
0.15	91.17%	9.9%	93.1%	41	76
0.10	90.97%	10.4%	93.3%	41	76
0.05	90.76%	10.7%	93.1%	41	76

5.3.2 "Improved" Ensemble Learning

In an attempt to increase the accuracy of the ensemble learning, the confidence factors of 0.20 and 0.15 are used on the C4.5 decision trees with boosting and bagging methods. Despite the increase in accuracy that is seen when using the confidence factor 0.20 and 0.15, the ensemble methods had a decrease in accuracy in comparison to the default confidence factor. This is likely due to the decrease in size of the various trees created by when performing the bagging and boosting methods. The larger confidence factor produces the better accuracy for both the boosting and bagging methods.

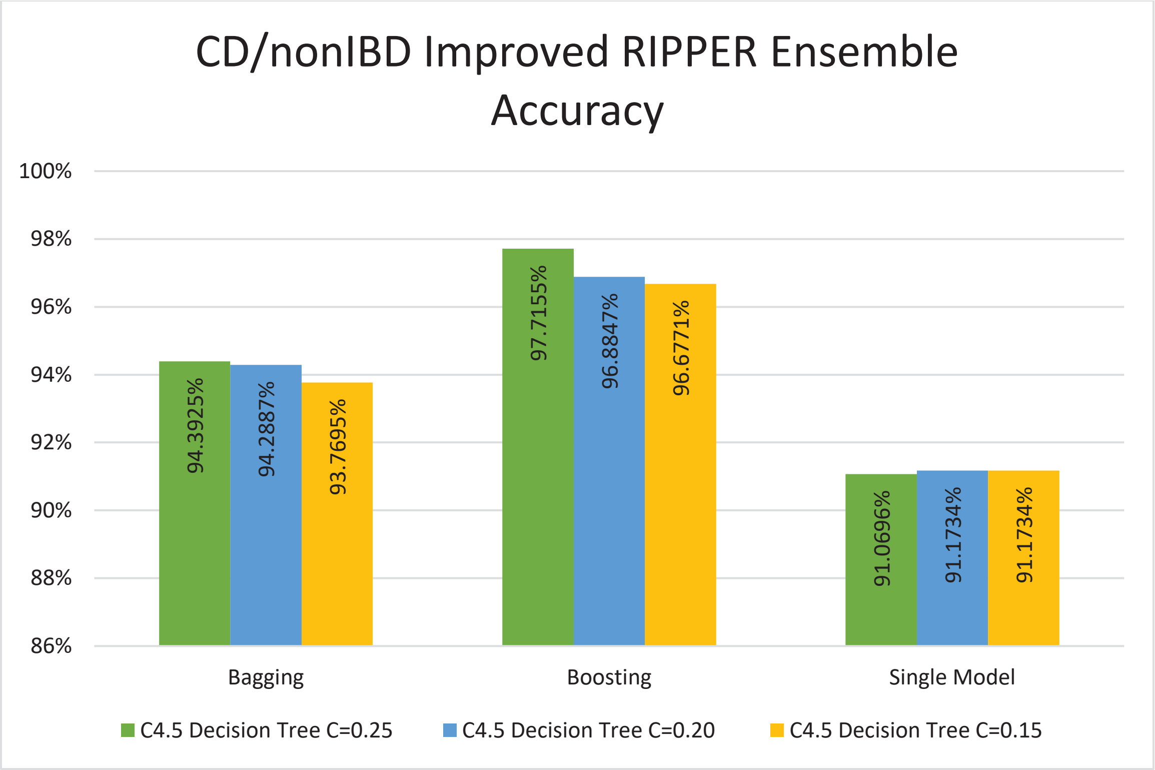

Figure 6: CD/nonIBD “Improved” C4.5 Ensemble Accuracy

Chapter 6. Binary Classification: UC and nonIBD

This section discusses the binary classification of instances as either UC or nonIBD. This binary classification focus on whether a patient has UC, so future tests would have to be performed to see if a patient has CD if the test comes up negative. This section will compare accuracies against those obtained by Geurt’s study but keep in mind that their study is comparing IBD and nonIBD [13]. A comparison is not been made to Mossotto’s study because the diagnosis is only between CD and UC without healthy controls [12].

6.1 Initial Results

Figure 5 shows a comparison between the different models on the different subsets on the UC/nonIBD dataset. The dataset consists of 374 instances of UC and 364 instances of nonIBD.

6.1.1 Decision Table

At first glance, it appears that the decision table using the *all* subset is the best model option for the CD and nonIBD model. The decision table has good readability and has the best accuracy of the models and subsets, but the rules created show that it is not a good option for this dataset. The table created with the *all* use following attributes: “Starch (white rice bread pizza potatoes yams cereals pancakes etc.) ”, “General wellbeing”, “Bowel frequency during the day”, “*Eggerthella*”, “*Lactobacillus*” and “*Haemophilus pittmaniae*”. Generally a smaller decision table usually means a better decision table, but asking for a patients age, whether they have had bowel surgery, their general wellbeing, and looking at concentration of one bacteria, the is not going to be enough information to diagnosis someone with a complex disease like IBD. The decision table using the *metadata* subset also has quite high accuracy, but it uses the same attributes as the *all* subset minus the " *Bacteroides dorei*" attribute. The decision table has more reasonable attribute selection for the other subsets, however, these tables are not as attractive as a choice because they have lower accuracy than the models with similar accuracy.

6.1.2 K-Nearest Neighbor

After the decision table, k-nearest neighbor has high accuracy in comparison to other models. The k-nearest neighbor with all the different subsets of this dataset also outperformed the k-nearest neighbor from Geurts's study which has an error rate of 20.21% (or accuracy of 79.79%) [13]. While the accuracy of the model is the highest, the model has low readability. Due to the difficulty of reading the k-nearest neighbor, no addition information can be gathered from the model which in turn means fewer medical professionals would trust the model's diagnosis.

6.1.3 Naïve Bayes

The Naïve Bayes model does not perform very well for all the different subsets with the exception of the *metadata* subset. This model appears to not be equipped to work with the metagenomic abundance data because the accuracy of this model increases with the two subsets that include the metadata (*all* and *metadata* subsets). However, the specific metadata attributes of sex and race appear to have no effect on the model since the accuracy of the *metagenomic*, *sex*, and *race* subsets are the same. Despite the Naïve

Bayes model being popular for medical diagnosis problems, it is not a good fit for this specific problem.

6.1.4 C4.5 Decision Tree

The C4.5 decision tree has the best accuracy when classifying UC and nonIBD when looking at the *all* subset. After the decision table and k-nearest neighbor models, C4.5 decision tree has the highest accuracy but only when using the *all* subset.
The C4.5 decision tree is quite consistent across the different subsets with the exception of the *metadata* subset, since the accuracy between subset has less than a percent difference. The good readability of the decision tree also makes this model more trusted because the user knows why the model has made its prediction. The C4.5 decision tree using the *all* outperforms the single tree in Geurts's study by about 17.74% [13], and it also outperforms a colonoscopy by 2.07% [10]. However, the accuracy of the best method in Geurts' study which is using attribute selection by boosting (r = 1%) outperformed the C4.5 decision tree but only by about 2.25% [13].

6.1.5 RIPPER

The RIPPER algorithm also has merit with the classification between IBD and nonIBD. When using the *all* subset, the accuracy is the highest after the decision table and k-nearest neighbor. Since the RIPPER is a rule-based classifier, the readability is excellent even for individual with little to no experience with the model itself. The size of the RIPPER algorithm increases its attractiveness in comparison with the C4.5 decision tree. The RIPPER algorithm has 9 rules while the C4.5 decision tree has a size of 42 with 22 leaves. However, the RIPPER algorithm only outperforms the colonoscopy by about 0.58% [10].

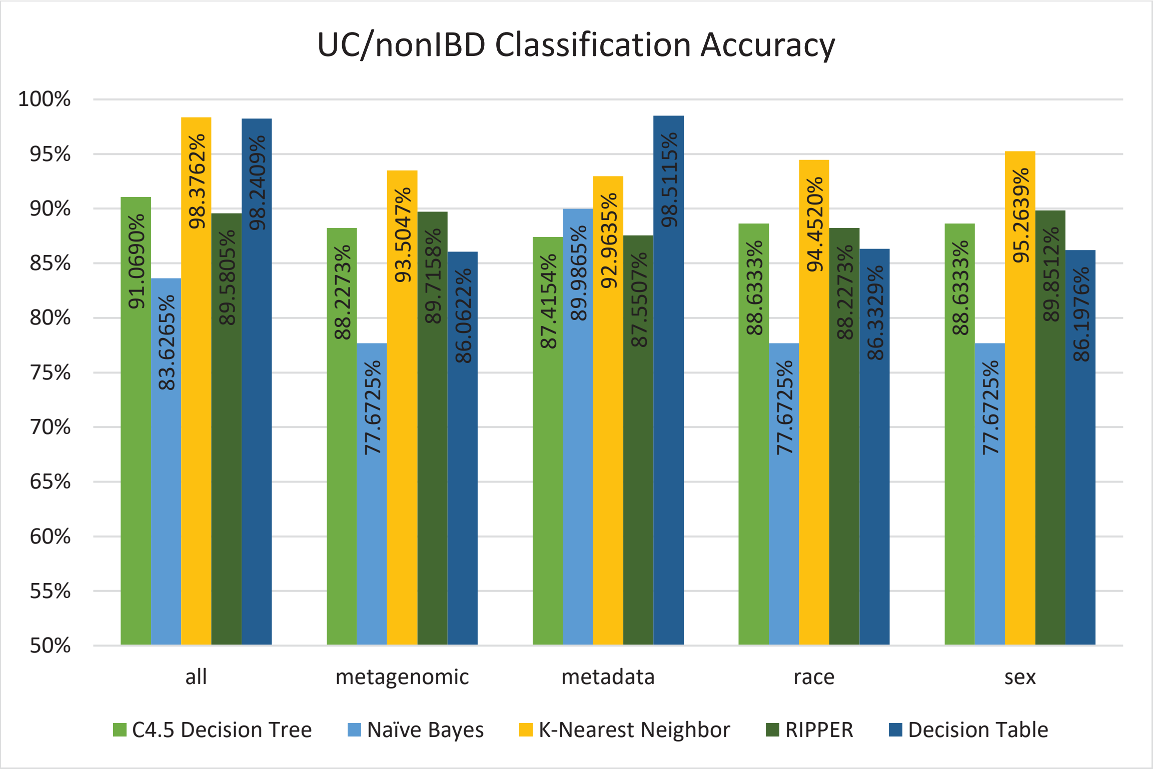

Figure 7: UC/nonIBD Classification Accuracy

6.2 Ensemble Learning

The ensemble methods improved the accuracy of both the C4.5 decision tree and the RIPPER algorithm. The boosting method performs better than bagging method. The RIPPER has a larger accuracy than the C4.5 decision tree for both the boosting and bagging method. In the bagging method, the decision tree has an increase of about 4.74% while the RIPPER algorithm of about 9.07%. The boosting, however, has a larger impact on the RIPPER algorithm's accuracy then the accuracy for the decision tree (with about 9.47% and 6.90% respectively). This study's bagging and boosting for trees outperform Geurt's study by about 10.64% and outperforms the boosting by about 9.18% [13].

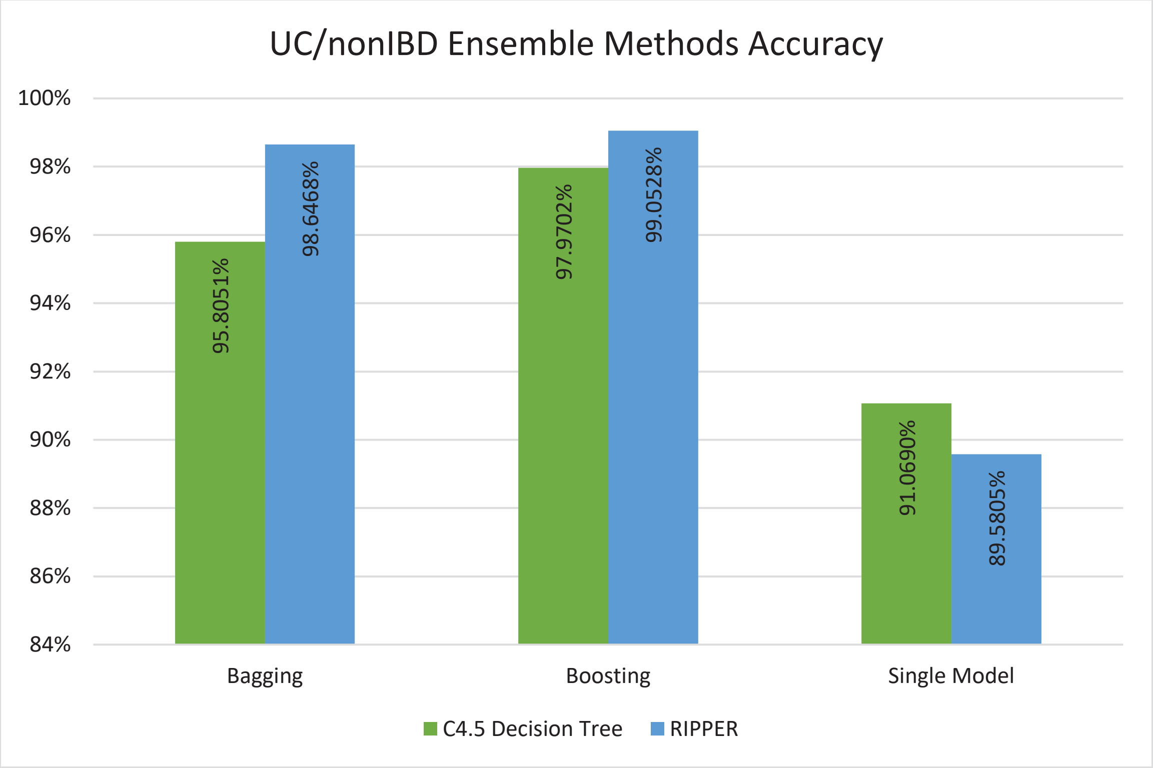

Figure 8: UC/nonIBD Ensemble Methods Accuracy

6.3 Accuracy Improvement

In an attempt to increase the accuracy of the C4.5 decision tree, the confidence factor related to pruning the tree has been changed. The default value in WEKA for the confidence factor is 0.25. The decrease of this number increases the amount of pruning. In this experiment, the confidence factor will only be lowered to guard against overfitting.

6.3.1 Best Model – C4.5 Decision Tree

There is no change in accuracy, false positive rate, size, or number of leaves between confidence factors of 0.25, 0.20, and 0.15, however, 0.20 and 0.15 have slightly increased ROC Area (by 0.01%). There is a decrease in accuracy by about 0.41%, a decrease in the ROC Area to 91.4%, and an increase in the false positive rate by 0.4%

when using the confidence factor 0.10. The confidence factor 0.05 provides a slightly better tree then using 0.15 with an increase of about 0.13% in accuracy and 0.4% in ROC Area and decrease in the false positive rate (by about 0.1%), the size, and number of leaves.

Table 4: C4.5 Decision Tree Optimization for UC/nonIBD Classification

Confidence Factor	**Accuracy**	**FP Rate**	**ROC Area**	**Number of Leaves**	**Size**
0.25	91.07 %	8.9%	91.8%	22	42
0.20	91.07 %	8.9%	91.9%	22	42
0.15	91.07 %	8.9%	91.9%	22	42
0.10	90.66%	9.3%	91.4%	22	42
0.05	90.79%	9.2%	91.8%	20	38

6.3.2 “Improved” Ensemble Learning

The confidence factors 0.20 and 0.15 were used to attempt to increase the accuracy of the ensemble learning methods. There was no increase in accuracy seen by the boosting or bagging methods when using either 0.20 or 0.15. Boosting sees no change in accuracy between using confidence factor 0.25 and 0.15 but there is a decrease in accuracy by about 0.54%. Bagging sees a decrease in accuracy of about 0.14% when using 0.20 and a decrease of about 0.41% when using 0.15.

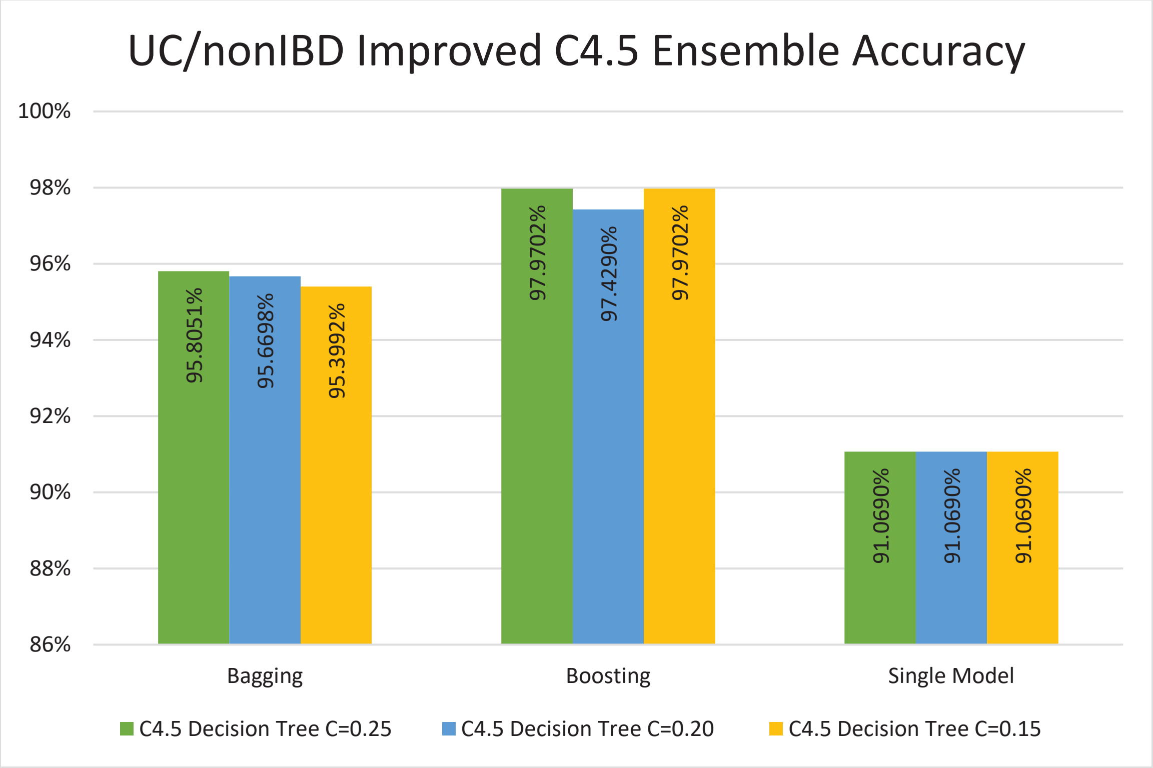

Figure 9: UC/nonIBD "Improved" C4.5 Ensemble Accuracy

Chapter 7. Ternary Classification: CD, UC, and nonIBD

This section discusses the ternary classification od instances as CD, UC or nonIBD. While the accuracy of the prediction does decreases but a ternary classification allows for the model give a more specific diagnosis between the specific types for IDB. In this section a comparison between the study and the study performed by Mossotto and the study performed by Geurts. Keep in mind that both studies are dealing with binary classifications and that Geurts' study is IBD vs nonIBD while the Mossotto's study is CD vs UC [12] [13].

7.1 Initial Results

Figure 6 shows a comparison between the different models on the different subsets on the ternary dataset. The dataset consists of 599 instances of CD, 375 instances of UC, and 364 instances of nonIBD.

7.1.1 Decision Table

At first glance, it appears when using the *all* subset is the best model option for the ternary classification model, but just like before the rules created show that it is not a good option for this dataset. The table created with the *all* use following attributes: "consent_age", "General wellbeing", "Bowel frequency during the day", "*Bacteroides fluxus*", and "*Bacteroides sp 4 3 47FAA*". Generally a smaller decision table usually means a better decision table, but asking for a patients age, their general wellbeing, the frequency of bowel movements during day and looking at concentration of two bacteria, the is not going to be enough information to diagnosis someone with a complex disease like IBD. The decision table using the *metadata* subset also has quite high accuracy, but it uses the following attributes: “In the past 2 weeks have you had diarrhea?”, “General wellbeing”, “Arthralgias”, and “Bowel frequency during the day”. These attributes are also not enough to diagnosis a complex disease. The decision table has more reasonable attribute selection for the other subsets, however, these tables are not as attractive as a choice because they have lower accuracy than the models with similar accuracy.

7.1.2 K-Nearest Neighbor

The k-nearest neighbor has the second highest accuracy after the decision table. Due to the poor readiability of the model it is hard to trust with a task such as medical diagnosis especially when there are other models with signicantly high accuracy that are easily readiable. The k-nearest neighbor from all the subsets outperformed the k-nearest neighbor from Geurts’s study. The *all* subset outperformed Geurts’ study by about 16.77% [13].

7.1.3 Naïve Bayes

The Naïve Bayes model does not perform very well for all the different subsets with the exception of the *metadata* subset. This model appears to not be equipped to work with the metagenomic abundance data because the accuracy of this model increases with the two subsets that include the metadata (*all* and *metadata* subsets). However, the specific metadata attributes of sex and race appear to have no effect on the model since the accuracy of the *metagenomic*, *sex*, and *race* subsets are the same. Despite the Naïve

Bayes model being popular for medical diagnosis problems, it is not a good fit for this specific problem.

7.1.4 RIPPER

The RIPPER algorithm has the best accuracy for ternary classification when using the *all* after the decision table and k-nearest neighbor. The RIPPER algorithm continues to provide a small resulting model than the C4.5 decision with the RIPPER algorithm creating 11 rules and the C4.5 decision tree with the best accuracy (using the *race* subset) has a size of 168 with 89 leaves. The RIPPER algoritm outperformed the optimizied SVM from Mossotto's study be about 7.7% [12] and also outperformed colonoscopy's accurracy by about 3.6% [10]. However, the accuracy of the best method in Geurts' study which is using attribute selection by boosting (r = 1%) outperformed the C4.5 decision tree but only by about 0.72% [13].

7.1.5 C4.5 Decision Tree

While the C4.5 decision tree is not the best option for the ternary classification problem, it still has significant merit. The accuracy of the decision tree surpasses the RIPPER algorithm when using the *metagenomic*, *race*, and *sex* subsets with the best accuracy using the *race* subset. The readability of the C4.5 decision tree still makes it an attractive model for a diagnosis problem; however, the larger size and the lower accuracy means it is still not a good as a choose as the RIPPER algorithm. The RIPPER algorithm is still the better of the two models but the C4.5 decision tree is still an attractive model that should still be investigated further, especially because of itself result when looking at the CD/nonIBD and the UC/nonIBD classification problems. This study has better accuracy in comparison to Mossotto's study by about 10.47% for the single tree model and the Guerts' study by 7.15% [12].

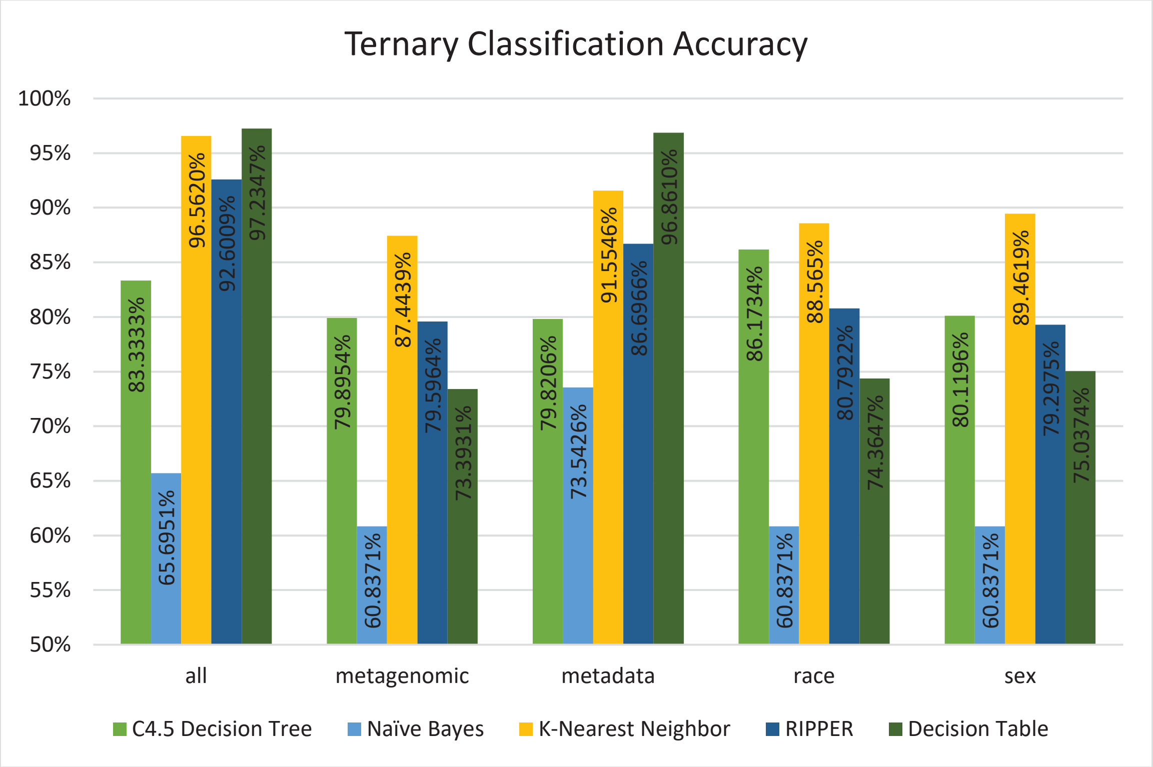

Figure 10: Ternary Classification Accuracy

7.2 Ensemble Learning

The ensemble methods improved the accuracy of both the C4.5 decision tree and the RIPPER algorithm when using the *all* subset. The bagging model increases the accuracy of the single method with about 9.49% increase for the decision tree and about 2.99% increase for the RIPPER algorithm. The boosting method also has a larger effect on the decision tree with an increase in acurracy of about 12.33%, while boosting increased the accuracy of the RIPPER algorithm by about 4.56%. This study's bagging and boosting for trees outperform Mossotto's study who's bagging tree accuracy of 77.6% and boosting tree's accuracy of 74.8% [12] and Guert's study who's bagging tree with an error rate of 13.36% (or accuracy of 86.64%) and boosting tree's error rate of 10.44% (or accuracy of 89.56%) [13].

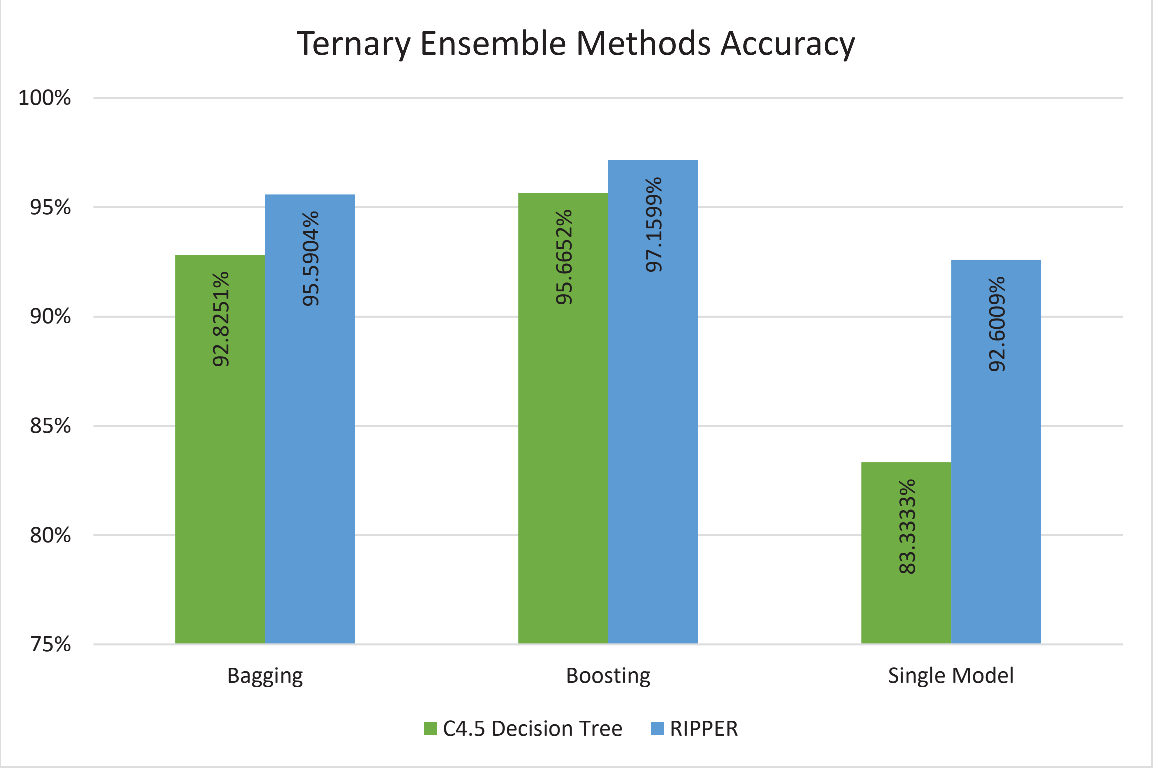

Figure 11: Ternary Ensemble Method Accuracy

7.3 Accuracy Improvement

In an attempt to improve the accuracy of the RIPPER algortihm, the number of optimzation runs performed during the optimization stage for both the single model and for ensemble learning. The default number of optimzation runs in WEKA is 2.

7.3.1 Best Model - RIPPER

For ternary classification, the increase of optimization runs generally increases the accuracy of the rule set until too many runs are performed. The number of rules created also have curve that peaks around 5 runs. Running 4 and 5 runs returns the accuracy but when performing the 5 runs, the false positive rate increase by 0.1% and number of rules created increases by 2 but the ROC Area increases by 0.1%. When 20 runs are peformed, number of rules does decrease but it also comes with a decrease in accuracy, an increase in false positive rate, and a decrease in ROC Area.

Table 4:RIPPER Optimization for Ternary Classification

Number of Optimizations Runs	Accuracy	FP Rate	ROC Area	Number of Rules
2	92.60%	4.5%	95.4%	14
3	92.75%	4.5%	95.5%	14
4	93.27%	4.2%	95.8%	15
5	93.27%	4.3%	95.9%	17
10	93.35%	4.2%	96.0%	16
20	92.83%	4.5%	95.2%	15

7.3.2 Improved Ensemble Learning

To continue improving accuracy, the RIPPER algortihm with 10 optizimation runs is used with the ensemble learning methods of bagging and boosting. As show above, bagging improves the 2 runs RIPPER algortihm by about 5.83% and boosting improves it by 7.40%. From the 2 run RIPPER alogrithm, there is an increase of about 2.54% when using bagging and an increase of about 4.86% when using boosting on the 20 run RIPPER algorithm. While the ensemble learning with the increased number of optizimation runs has a high accuracy, it has a smaller increase of the accuracy. The bagging method has very little change in accuracy with only around a 0.3% change between 2 runs and 10 runs.

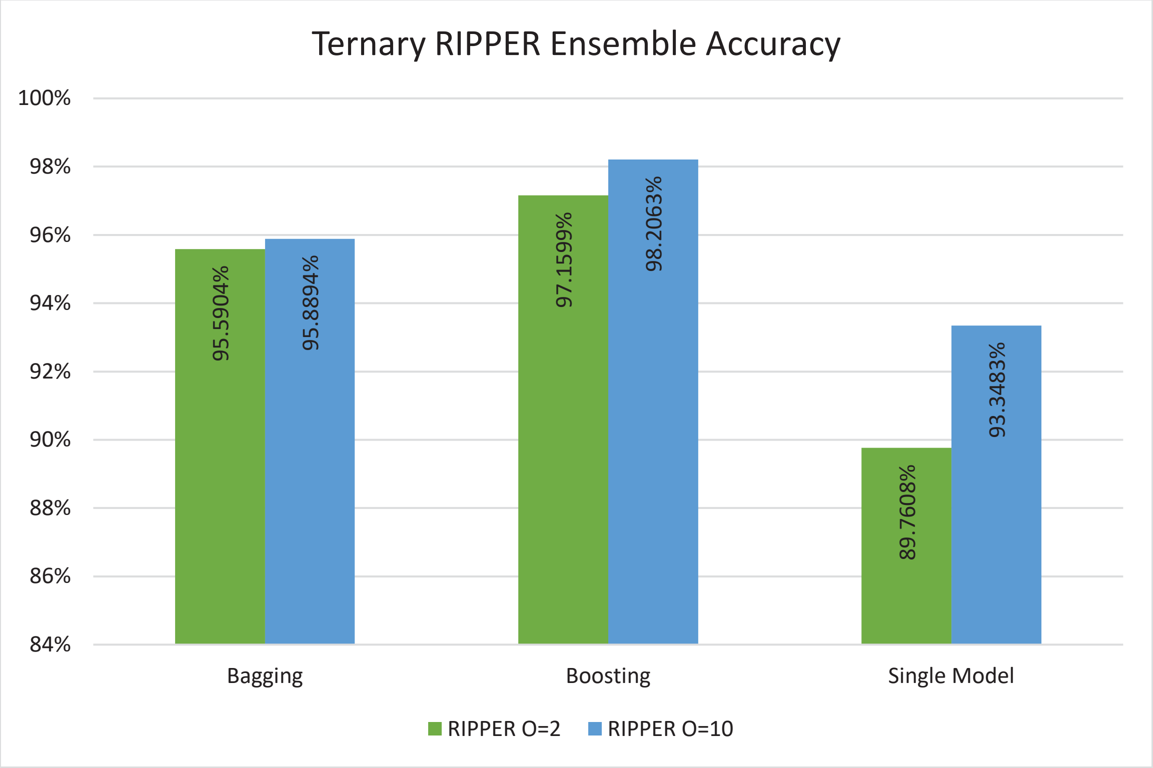

Figure 12: Ternary Improved RIPPER Ensemble Accuracy

Chapter 8. Discussion

8.1 Change in Bacteria Abundances

8.1.1 Genus Coprococcus

The root of the decision trees for all of the subset, exclude the *metadata* subset, *Coprococcus eutactus* for both IBD/nonIBD and ternary classification. While *Coprococcus eutactus* is not the root of the CD/nonIBD tree, it is located on the second level. *Coprococcus eutactus* is not found in the decision tree created for UC/nonIBD. The bacterica *Coprococcus comes* is found in the rules created by the RIPPER algorithm for the IBD/nonIBD problem. The placement of this species of bacteria makes senses given knowledge about its genus. Studies show that the abundance of genus *Coprococcus* decrease when a patient has IBD [25] [26]. When C*oprococcus eutactus* is great than zero percent of the total microbiota the vast majority of the instances on that side of the tree are diagnosis nonIBD. While the genus *Coprococcus* has a decrease in abundance for IBD patients, my research shows that it may have a more important role in CD then UC. *Coprococcus* is also found to help protect patients against colon cancer which IBD patients are at a higher risk of getting [27].

8.1.2 Genus Alistipes

It also appears that the bacteria genus *Alistipes* is important in the classification of IBD. All of the rules created by the RIPPER algorithm using the genus A*listipes* classify as nonIBD if the abundance of bacteria is higher than a specific point. Multiple bacteria species in this genus also appear in the varies decision trees created. This genus appears in at least on rule creates by the RIPPER algorithm for all four of the classification problems. The genus *Alistipes* appears in the trees created by all the classification problem except UC/nonIBD, with it appearing relatively high in the IBD/nonIBD tree. This would indicate that this genus may have a stronger link the CD then to UC. This pattern follows because multiple studies so that multiple bacteria species from this genus (*alistipes finegoldii, alistipes putredinis,* and *alistipes shahii)* decrease in amount and/or growth rate in people with IBD in comparison to healthy controls [28] [29]. All of the previously mention bacteria species have been specifically seen as a part of some of model created.

8.1.3 Genus Bacteroides

The bacteria genus *Bacteroides* is the genus that is found most often seen in the rules created by the RIPPER algorithm and the nodes of the C4.5 decision trees across all four classification problems. This observation does make sense because this genus has been shown to have a strong link to IBD [30]. Unlike the previous bacteria genera discussed, there is not a set pattern seen in the rule for an increase or decrease of abundance for any class in particular. This may be caused by different species in the genus acting in different ways then each other. However, in a meta-analysis performed by Yingting Zhou and Fachao Zhi, it was found that different studies have shown contradictory results when looking at genus *Bacteroides* and IBD. In the FISH study and conventional culture studies they looked at, CD and UC patient are shown to higher levels of *Bacteroides* then healthy controls, but in the Real-Time Quantitative PCR studies, CD and UC patient are shown to lower levels of *Bacteroides* then healthy controls [30]. Yingting Zhou and Fachao Zhi theorized that this difference may be caused difference in ethnicity because the FISH study and conventional culture studies received samples from mostly European patient while the Real-Time Quantitative PCR studies were mostly Asian patients. This may explain the variety of levels seen in this study with a wider variety of different ethnic groups.

8.1.4 Specific Species

Other bacteria that are known to decrease in abundance with patient diagnosed with IBD are *Clostridium leptum* and *Faecalibacterium prausnitzi* [31]. They are both present in the rules created by the RIPPER algorithm for ternary classification (with the *all* subset) and when the abundance is greater than or equal to a specific point then they are classified as nonIBD. This follows from a decrease seen in IBD patients. However, in the C4.5 decision tree for ternary classification (with the *all* subset), *Clostridium leptum* is not present at all and the relation of *Faecalibacterium prausnitzi* between IBD patient and nonIBD is flipped. When the abundance is higher than a specific point then they are classified as UC and nonIBD when less than or equal to that specific point.

8.2 Race and Ternary Classification

As previously stated, the ethnicity/race of a person can increase the chances of IBD [6]. However, race seems to play a bigger role when diagnosis on a ternary classification then IBD/nonIBD, CD/nonIBD, and UC/nonIBD classification. The best accuracy for the C4.5 decision tree on the ternary classification is when using the *race* subset. It increases the accuracy by about 3 percent when using the *all* subset and about 6 percent when using the *metagenomic* subset. This increase can also be seen in the RIPPER algorithm. While the accuracy for RIPPER using the *race* subset is not larger than the accuracy of the *all* subset, there is little under a 3 percent increase in comparison to the *metagenomic* subset. In contrast, the accuracy between the *all*, *metagenomic* and *race* do not even differ by a single percent when using the C4.5 tree with *metagenomic* subset having the best accuracy. The RIPPER algorithm has more variability with the *all* subset have about 2 percent higher than either the *metagenomic* or *race* subsets. The *metagenomic* and *race* subsets differ less than a percent. This relation to the *race* attribute is not seen in the UC/nonIBD problem and can only be seen in the CD/nonIBD problem in the accuracies of the RIPPER algorithm (by about 0.83%). This leads one to believe that the attribute race has more importance to distinguishing between CD and UC then diagnosing IBD itself or diagnosing one or the other individually, with slightly more importance to CD diagnosis.

8.3 Study Comparison

When using the RIPPER algorithm with the *all* subset consistently suppressed the accuracy of diagnoses from the other studies discussed in the related work. It is likely for two reasons. The first reason is the size of the dataset is significantly larger than in previous studies. The Metagenomic dataset used for this study is about 60% larger than the two related studies discussed earlier in this paper combined. Even the smallest classification (UC/nonIBD) is about 29% larger than the other two studies combined. With a larger dataset, the models created will be better trained models. The second reason is the type of data. While the protein profiles generated by the mass spectrometry, does provide a diagnose data for a disease, metagenomic data from a stool sample provides information specific to IBD and patient’s microbiota. This shifts the focus specifically on the abundancies of bacteria in a patient’s GI tract provides insight specifically where IBD affects. Of the these two reason is the more significant of the two reasons is the larger size of the dataset.

This study also allows for the investigation of ternary classification of IBD. The previous related studies only looks at binary classification (either IBD vs nonIBD or CD vs UC). This study also investigates multiple binary classification problems (IBD/nonIBD, CD/nonIBD, and UC/nonIBD). Looking at both of these two avenues provides a better look into the complexity of the larger diagnosis problem. The ternary classification continued to outperform the accuracy of the previous studies even with the added complexity with the addition to a third class. The discussion of the different classification problems in this study is investigated further in the next section.

8.4 Classification Type Comparison

As stated before, the ternary classification, in general, is going to have lower accuracy then that of the binary classification counterparts. One model, however, does not follow this tendency, which is the RIPPER algorithm. The IBD/nonIBD accuracy is similar to the Ternary RIPPER algorithm with the CD/nonIBD and UC/nonIBD trailing behind by about 1%. This section will draw a further comparison between IBD/nonIBD and Ternary classification and further comparison between CD/nonIBD and UC/nonIBD.

8.4.1 IBD/nonIBD vs Ternary

The RIPPER algorithm when using the *all* dataset has a very slight increase (by 0.1495% with two more instances properly classified) in its accuracy between IBD/nonIBD and ternary classification when using the default parameters. Not only is there an increase in accuracy but allows the precision (by 0.2%), recall (by 0.1%), F-measure (by 0.2%), MCC (by 7.7%), ROC area (by 4.9%), and (Precision Recall) PRC Area (by 0.6%) and a decrease in the false positive rate by 7.6%. This allows the ternary classification to be used for the RIPPER algorithm without losing accuracy but gaining more information for a more specific diagnosis between CD, UC, and nonIBD.

Table 5: RIPPER Performance Metrics for Different Optimization Runs

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
IBD/N O = 2	92.5%	12.1%	92.4%	92.5%	92.4%	80.8%	90.5%	90.6%
IBD/N O = 20	93.9%	8.5%	94.0%	93.9%	94.0%	84.8%	93.0%	92.6%
Ternary O = 2	92.6%	4.5%	92.8%	92.6%	92.6%	88.5%	95.4%	91.2%
Ternary O = 10	93.3%	4.2%	93.5%	93.3%	93.4%	89.6%	96.0%	92.3%

This increase in accuracy of the ternary classification does not continue to hold when the number of optimizations runs (O) is adjusted (Table 5). The binary classification accuracy increases and suppresses the ternary classification but only by a fraction of a percent. However, the ternary classification still improves upon the binary classification with cutting the false positive rate in half and with and increase in the MCC and the ROC Area. With the difference in accuracy only being less than a percentage point, the decrease of the false positive rate by 4.3% still makes the ternary classification RIPPER algorithm the superior choice.

This flip in accuracy does not continue for the bagging technique. The ternary classification accuracy for bagging trailing the binary classification by 1.1% when using only 2 optimization runs. When using their optimal number of optimization runs with the bagging method, binary classification beat ternary classification by about 1.4%. The false positive rates for the ternary classification continue to outperform that of the binary classification with the ternary improving by 4.5% when both using 2 runs and by 3% when using the optimal number of runs. We can also see a that MCC is higher for the ternary classification using bagging then binary by 1.4% when using 2 runs and by 0.3% when using the optimal number of runs.

Table 6: RIPPER Performance Metrics for Bagging

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**
IBD/N O = 2	96.7%	7.6%	96.7%	96.7%	96.7%	91.6%	99.3%	99.3%
IBD/N O = 20	97.3%	5.8%	97.3%	97.3%	97.3%	93.2%	99.4%	99.2%
Ternary O = 2	95.6%	3.1%	95.7%	95.6%	95.65%	93.0%	99.2%	98.9%
Ternary O = 10	95.9%	2.8%	96.0%	95.9%	93.9%	93.5%	99.3%	99.0%

This flip in accuracy does not continue for the boosting technique. The ternary classification accuracy for boosting trailing the binary classification by 1.3% when using only 2 optimization runs. When using their optimal number of optimization runs with the boosting method, binary classification beat ternary classification by about 1.1%. The false positive rates for the ternary classification continue to outperform that of the binary classification with the ternary improving by 1.2% when both using 2 runs but when using the optimal number of runs, both binary and ternary classification have the same false positive rate.

Table 7: RIPPER Performance Metrics for Boosting

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
IBD/NO = 2	98.5%	3.1%	98.5%	98.5%	98.5%	96.2%	99.8%	99.8%
IBD/NO = 20	99.3%	1.1%	99.3%	99.3%	99.3%	98.3%	99.9%	99.9%
Ternary O = 2	97.2%	1.9%	97.2%	97.2%	97.2%	95.6%	99.6%	99.3%
Ternary O = 10	98.2%	1.1%	98.2%	98.2%	98.2%	97.2%	99.7%	99.6%

Despite the binary classification methods having better accuracy when performing more optimization runs and when using ensemble learning methods, the ternary classification is still the better option. The ternary classification also for a more specific diagnosis and has consistently provide a lower false positive rate, except for using boosting on the optimal number of runs. While using the binary classification using 20 optimization runs with boosting would provide the highest accuracy with a low false positive rate, it loses the ease of readability that is provided by the single model and, as previously said, IBD/nonIBD classification does not provide the specific diagnosis that ternary classification provides.

8.4.2 CD/nonIBD vs UC/nonIBD

The CD/nonIBD problem has a small increase in accuracy that UC/nonIBD does not see when the confidence factor decreases. The only change seen between confidence factor 0.25, 0.20, and 0.15 is an increase of 0.01% in the ROC Area and PRC Area with 0.20 and 0.15. Despite the slightly better accuracy seen in the CD/nonIBD problem (0.1%) the UC/nonIBD consistently see a better false positive rate with a decrease of 1% from the best CD/nonIBD tree and a better MCC with an increase of 0.9%. There is a small fluctuation in the ROC Area when looking at the change between 0.25, 0.20, and 0.15 in the CD/nonIBD. There is a drop of 0.1% between 0.25 and 0.20 and gained that 0.1% back when using 0.15.

Table 8: C4.5 Performance Metrics using Different Confidence Factors

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**
CD/N C=0.25	91.1%	10.2%	91.1%	91.1%	91.1%	81.0%	93.1%	91.7%
CD/N C=0.20	91.2%	10.0%	91.2%	91.2%	91.2%	81.2%	93.0%	91.6%
CD/N C=0.15	91.2%	9.9%	91.2%	91.2%	91.2%	81.2%	93.1%	91.8%
UC/N C=0.25	91.1%	8.9%	91.1%	91.1%	91.1%	82.1%	91.8%	89.4%
UC/N C=0.20	91.1%	8.9%	91.1%	91.1%	91.1%	82.1%	91.9%	89.5%
UC/N C=0.15	91.1%	8.9%	91.1%	91.1%	91.1%	82.1%	91.9%	89.5%

The bagging method does not show much potential for the CD/nonIBD and UC/nonIBD with the smaller confidence factors. When investigating the bagging method, the adjusted confidence factors for both CD/nonIBD and UC/nonIBD does not provide promising results with none of the metrics improved. However, UC/nonIBD had a larger increase in accuracy between the single tree and the bagging method than the CD/nonIBD despite its larger accuracy in the single trees. UC/nonIBD also continues to have a lower false positive rate than CD/nonIBD.

Table 9: C4.5 Performance Metrics for Bagging

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
CD/N C=0.25	94.4%	7.2%	94.4%	94.4%	94.4%	88.0%	98.7%	98.7%
CD/N C=0.20	94.3%	7.4%	94.3%	94.3%	94.3%	87.8%	98.6%	98.7%
CD/N C=0.15	93.8%	8.0%	93.8%	93.8%	93.7%	86.7%	98.6%	98.6%
UC/N C=0.25	95.8%	4.2%	95.8%	95.8%	95.8%	91.6%	98.9%	98.9%
UC/N C=0.20	95.7%	4.3%	95.7%	95.7%	95.7%	91.4%	98.9%	98.8%
UC/N C=0.15	95.4%	4.6%	95.4%	95.4%	95.4%	90.8%	98.8%	98.8%

The bagging method does not show much potential for the CD/nonIBD and UC/nonIBD with the smaller confidence factors. Just like the bagging method, the adjusted confidence factors for both CD/nonIBD and UC/nonIBD does not provide promising results with none of the metrics improved. UC/nonIBD has better accuracy and false positive rate then CD/nonIBD just like is seen in the bagging method discussed before. There is a small fluctuation in the accuracy (TP Rate) when looking at the change between 0.25, 0.20, and 0.15 in the CD/nonIBD. There is a drop of 0.6% between 0.25 and 0.20 and gained that 0.1% back when using 0.15.

Table 10: C4.5 Performance Metrics for Boosting

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
CD/N C=0.25	97.7%	3.1%	97.7%	97.7%	97.7%	95.1%	99.6%	99.7%
CD/N C=0.20	96.9%	3.7%	96.9%	96.9%	96.9%	93.4%	99.4%	99.4%
CD/N C=0.15	96.7%	4.3%	96.7%	96.7%	96.7%	92.9%	99.5%	99.6%
UC/N C=0.25	98.0%	2.0%	98.0%	98.0%	98.0%	95.9%	99.8%	99.8%
UC/N C=0.20	97.4%	2.6%	97.4%	97.4%	97.4%	94.9%	99.8%	99.8%
UC/N C=0.15	98.0%	2.0%	98.0%	98.0%	98.0%	95.9%	99.8%	99.8%

Looking CD/nonIBD and UC/nonIBD using this dataset does not seem to be an avenue for future study for this particular dataset. While they do provide accuracies that are improvements to other studies, they do not complete with the ternary classification. The ternary classification provides a diagnosis that is more specific because it can determine if a patient has CD, UC, or nonIBD with a model that has higher accuracy and a lower false positive rate.

8.5 Colonoscopy

The study produces many models that beat the accuracy of a colonoscopy. If one of these models or similar models could be incorporated or formed into medical diagnostic test, a stool test could be more accurate than an endoscopy like a colonoscopy.

Out of the current tests ordered to diagnosis, patients are most comfortable with a stool test. Patients have less anxiety about the pain, cost, outcome, and time of a stool test then a colonoscopy [9]. This would allow patients to only need to undergo a colonoscopy if it is truly necessary. If a diagnosis does come back as IBD positive, the patient would eventually need to have a colonoscopy perform, however this would allow for anyone without IBD to forgo the stress and expensive of a colonoscopy. The current stool test is also one of the test that patients are most comfortable with and is the least like to be refused by the patient when suggested by a medical professional [9]. This would allow medical professionals to have a test with higher accuracy that would be less likely refused by the patient and not increase the nerves of a patient without cause until more information can be collected.

8.6 Future Work

Future work will focus on attribute selection and importance. This would include making a data subset with only concentrations of bacteria genus and subset with only bacteria species concentrations, instead of looking at the concentrations of all the following: kingdom, phylum, class, order, family, genus, species, and trinomen (subspecies). This would determine if a bacteria genus has more or less of an impact on diagnosis then knowing the specific bacteria species. The only model that outperformed this study was using attribute selection by boosting (r = 1%) during the Geurts' study [13] so attribute selection has potential to improve the accuracy.

The investigation what attributes from the metadata are important to the diagnosis of the IBD. Also, the use of WEKA's attribute evaluator, such as CfsSubsetEval and

InfoGainAttributeEval, to help determine the importance of different data subsets and/or the importance of specific attributes to the problem.

Addition future work would also investigate the varies other datasets that are available on the IBDMDB. These datasets include 16S, Serology, Proteomics, Viromics, Metabolites, Metatranscriptomes, and Host Transriptomes datasets collected by the HMP2. This would allow for an investigation to what types of data is more useful for the diagnosis problem at hand.

Chapter 9. Conclusion

The RIPPER algorithm is the best choice for both IBD/nonIBD and Ternary classification. While there is a decision tree with a larger accuracy when looking at IBD/nonIBD classification, the size of the tree is significantly larger than the RIPPER algorithm which has accuracy that is roughly the same. The C4.5 decision tree provides a better model for CD/nonIBD and UC/nonIBD, however their accuracy cannot compete with the IBD/nonIBD and Ternary classification. The RIPPER algorithm using the Ternary classification with the *all* subset is the best models given this problem given its accuracy, readability, and its ability to provide a more specific diagnosis.

The decrease of bacteria genus *Alistipes*, genus *Coprococcus* and other bacteria species have been shown to increase the chance of IBD in patients. These bacteria are present and important in the C4.5 decision trees and the rules created by the RIPPER algorithm across multiple classification problems. While race/ethnicity is known to have an effect on a patient likelihood of having IBD, it appears to have a larger impact on the CD/nonIBD and Ternary classification with a more significant impact on Ternary classification.

The models presented in this study have promise in the diagnosis of IBD because their accuracy surpass that of the colonoscopy. It would also have the potential to provide a test that is less costly and less stressful for patients and less likely to be refused when ordered by a medical professional.

References

[1] S. C. Ng, H. Y. Shi, N. Hamidi, F. E. Underwood, W. Tang, E. I. Benchimol, R. Panaccione, S. Ghosh, J. C. Y. Wu, F. K. L. Chan, J. J. Y. Sung and G. G. Kaplan, "Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies," *The Lancet,* vol. 390, no. 10114, pp. 2769-2778, 2017.

[2] Centers for Disease Control and Prevention, "Inflammatory bowel disease (IBD): What is IBD?," 22 March 2018. [Online]. Available: https://www.cdc.gov/ibd/what-is-IBD.htm.

[3] T. M. Connelly and W. A. Koltun, "The surgical treatment of inflammatory bowel disease-associated dysplasia," *Expert Review of Gastroenterology & Hepatology,* vol. 7, no. 4, pp. 307-322, 2013.

[4] F. Xu, J. M. Dahlhamer, E. P. Zammitti, A. G. Wheaton and J. B. Croft, "Health-Risk Behaviors and Chronic Conditions Among Adults with Inflammatory Bowel Disease — United States, 2015 and 2016," *MMWR Morb Mortal Wkly Rep,* vol. 67, no. 6, pp. 190-195, 2018.

[5] Y. H. Kwon and Y. J. Kim, "Pre-diagnostic Clinical Presentations and Medical History Prior to the Diagnosis of Inflammatory Bowel Disease in Children," *Pediatr Gastroenterol Hepatol Nutr,* vol. 16, no. 3, pp. 178-184, 2013.

[6] M. P. C. Santos, C. Gomes and J. Torres, "Familial and ethnic risk in inflammatory bowel disease," *Ann Gastroenterol,* vol. 31, no. 1, pp. 14-23, 2018.

[7] Y. Ye, Z. Pang, W. Chen, S. Ju and C. Zhou, "The epidemiology and risk factors of inflammatory bowel disease," *Int J Clint Exp Med,* vol. 8, no. 12, pp. 22529-22542, 2015.

[8] I. Khan, N. Ullah, L. Zha, Y. Bai, A. Khan, T. Zhao, T. Che and C. Zhang, "Alteration of Gut Microbiota in Inflammatory Bowel Disease (IBD): Cause or Consequence? IBD Treatment Targeting the Gut Microbiome," *Pathogens,* vol. 8, no. 3, 2019.

[9] I. Noiseux, S. Veilleux, A. Bitton, R. Kohen, L. Vachon, B. W. Guay and J. D. Rioux, "Inflammatory bowel disease patient perceptions of diagnostic and monitoring tests and procedures," *BMC Gastroenterol,* vol. 19, no. 30, 2019.

[10] M. A. T. Passos, F. C. Chaves and N. Chaves-Junior, "The Importance of Colonoscopy in Inflammatory Bowel Diseases," *Arq Bras Cir Dig,* vol. 31, no. 2, 2018.

[11] E. Parva, R. Boostani, Z. Ghahramani and S. Paydar, "The Necessity of Data Mining in Clinical Emergency Medicine; A Narrative Review of the Current Literatrue," *Bull Emerg Trauma,* vol. 5, no. 2, pp. 90-95, 2017.

[12] E. Mossotto, J. J. Ashton, T. Coelho, R. M. Beattie, B. D. MacArthur and S. Ennis, "Classification of Paediatric Inflammatory Bowel Disease using Machine Learning," *Sci Rep,* vol. 7, no. 2427, 2017.

[13] P. Geurts, M. Fillet, D. de Seny, M.-A. Meuwis, M. Malaise, M.-P. Merville and L. Wehenkel, "Proteomic mass spectra classification using decision tree based ensemble methods," *Bioinformatics,* vol. 21, no. 14, pp. 3138-3145, 2005.

[14] M. G. Attalla, S. B. Singh, R. Khalid, M. Umair and E. Epenge, "Relationship between Ulcerative Colitis and Rheumatoid Arthritis: A Review," *Cureus,* vol. 11, no. 9, 2019.

[15] F. Gigliucci, F. A. B. von Meijenfeldt, A. Knijn, V. Michelacci, G. Scavia, F. Minelli, B. E. Dutilh, H. M. Ahmad, G. C. Raangs, A. W. Friedrich, J. W. A. Rossen and S. Morabito, "Metagenomic Characterization of the Human Intestinal Microbiota in Fecal Samples from STEC-Infected Patients," *Front Cell Infect Microbiol,* vol. 8, no. 25, 2018.

[16] NIH Human Microbiome Project, "The Inflammatory Bowel Disease Multi'omics Database," [Online]. Available: https://ibdmdb.org.

[17] J. Lloyd-Price, C. Arze, A. N. Ananthakrishna, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn, D. Casero, H. Courtney, A. Gonzalez, T. G. Graeber, A. B. Hall, K. Lake, C. J. Landers, H. Mallick, D. R. Plichta, M. Prasad, G. Rahnavard, J. Sauk, D. Shungin, Y. Vázquez-Baeza, R. A. White III., J. Bishai, K. Bullock, A. Deik, C. Dennis, J. L. Kaplan, H. Khalili, L. J. McIver, C. J. Morgan, L. Nguyen, K. A. Pierce, R. Schwager, A. Sirota-Madi, B. W. Stevens, W. Tan, J. J. ten Hoeve, G. Weingart, R. G. Wilson, V. Yajnik, J. Braun, L. A. Denson, J. K. Jansson, R. Knight, S. Kugathasan, D. P. B. McGovern, J. F. Petrosino, T. S. Stappenback, W. S. Harland, C. B. Clish, E. A. Franzosa, H. Vlamakis, R. J. Xavier and C. Huttenhower, "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases," *Nature,* vol. 569, no. 7758, pp. 655-662, 2019.

[18] T. R. Patil and S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," *International Journal Of Computer Science And Applications,* vol. 6, no. 2, pp. 256-261, 2013.

[19] W. W. Cohen, "Fast Effective Rule Induction," *Machine Learning: Proceedings of the Twelfth International Conference,* pp. 115-123, 1995.

[20] R. Kohavi, "The Power of Decision Tables," *8th European Conference on Machine Learning,* pp. 174-189, 1995.

[21] M. Langarizadeh and F. Moghbeli, "Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review," *Acta Inform Med,* vol. 24, no. 5, pp. 364-369, 2016.

[22] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers," *Mult Classif Syst,* 2007.

[23] L. Breiman, "Bagging Predictors," *Machine Learning,* vol. 24, pp. 123-140, 1996.

[24] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," *Thirteenth International Conference on Machine Learning,* pp. 148-156, 1996.

[25] K. Nishino, A. Nishida, R. Inoue, Y. Kawada, M. Ohno, S. Sakai, O. Inatomi, S. Bamba, M. Sugimoto, M. Kawahara, Y. Naito and A. Andoh, "Analysis of endoscopic brush samples identified mucosa-associated dysbiosis in inflammatory bowel disease," *Journal of Gastroenterology,* vol. 53, no. 1, pp. 95-106, 2018.

[26] H. Nagao-Kitamoto and N. Kamada, "Host-microbial Cross-talk in Inflammatory Bowel Disease," *Immune Netw,* vol. 17, no. 1, pp. 1-12, 2017.

[27] D. Ai, H. Pan, X. Li, Y. Gao, G. Liu and L. C. Xia, "Identifying Gut Microbiota Associated With Colorectal Cancer Using a Zero-Inflated Lognormal Model," *Front Microbiol,* vol. 10, no. 826, 2019.

[28] T. G. J. de Meij, E. F. J. de Groot, C. F. W. Peeters, N. K. H. de Boer, C. M. F. Kneepkens, A. Eck, M. A. Benninga, P. H. M. Saveikoul, A. A. van Bodergraven and A. E. Budding, "Variability of core microbiota in newly diagnosed treatment-naïve paediatric inflammatory bowel disease patients," *PLos One,* vol. 13, no. 8, 2018.

[29] Q. He, Y. Gao, Z. Jie, X. Yu, J. M. Laursen, L. Xiao, Y. Li, L. Li, F. Zhang, Q. Feng, X. Li, J. Yu, C. Liu, P. Lan, T. Yan, X. Liu, X. Xu, H. Yang, J. Wang, L. Madsen, S. Brix, J. Wang, K. Kristiansen and H. Jia, "Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients," *Gigascience,* vol. 6, no. 7, pp. 1-11, 2017.

[30] Y. Zhou and F. Zhi, "Lower Level of Bacteroides in the Gut Microbiota Is Associated with Inflammatory Bowel Disease: A Meta-Analysis," *Biomed Res Int.,* 2016.

[31] S. Vrakas, K. C. Mountzouris, G. Michalopoulos, G. Karamanolis, G. Papatheodoridis, C. Tzathas and M. Gazouli, "Intestinal Bacteria Composition and Translocation of Bacteria in Inflammatory Bowel Disease," *PLoS One,* vol. 12, no. 1, 2017.