LITERATURE REVIEW: A Comparative Analysis of Machine Learning Methods for Predicting the Presence of Crohn's Disease

Aatreyi Pranavbhai Mehta
School of Computer Science
Carleton University
Ottawa, Canada K1S 5B6
aatreyipranavbhaime@cmail.carleton.ca

December 2, 2021

1 Introduction

Crohn's disease(CD) is an inflammatory bowel disease that is chronic and relapsing. It is distinguished by a transmural granulomatous inflammation that can affect any part of the gastrointestinal tract, most commonly the ileum, colon, or both. Its prevalence has continued to increase over the last 50 years, with northern Europe, the United Kingdom, and North America having the highest occurrence[6]. Despite the fact that biological treatment is associated with improved health-related quality of life, patients report difficulties in adjusting to lifestyle and daily activities during flares and remissions. Parallel Computing helped in enabling the study of gut microbiome. Trillions of microbes can be found primarily in the intestine and on skin. The majority of microbes in the intestines are found in the cecum, a "pocket" of the large intestine, and are referred to as the gut microbiome. The human gut microbiome contains up to 1,000 kinds of bacteria, each of which serves a unique role in your body. The majority are beneficial to your health, but some may cause illness. Researchers discovered that people with Crohn's disease had an excess of a type of gut bacteria known as adherent-invasive Escherichia coli (AIEC), which increases intestine inflammation[1]. Their research demonstrated that a bacteria-produced metabolite interacts with immune system cells in the intestinal lining, causing inflammation. In this way, parallel computing helped to lead to the discovery of the type of bacteria that causes Crohn's disease. The screening of faecal calprotectin (FC) levels has been found to be a therapeutically helpful method of monitoring CD development. FC is the most abundant protein in neutrophils and is a dimeric calcium, iron, manganese, and zinc sequestering protein. Increased FC concentrations in patients' faeces are a reliable indicator of IBD, making them ideal for monitoring disease development in CD patients and lowering the frequency of needless endoscopies. Machine learning has been used widely in the previous decade to predict disease progression, including a recent study that identified FC as a substantial risk factor for CD development. The best machine learning model is not always obvious, and comparison evaluations are required to determine the model with the highest predictive performance. In this paper, I aim to provide a comparative analysis between different machine learning methods based on their efficiency to predict the presence of CD given a huge

data set of patients.

2 Literature Review

A colonoscopy is a procedure that examines the large intestine and rectum for changes or abnormalities. A long, flexible tube is introduced into the rectum during a colonoscopy so that the doctor can examine the inside of the colon with the help of a tiny video camera at the tube's tip. One of the prior paper[7] suggested the preference of patients for colonoscopy and CT examinations (X-ray computed tomography) over a comparatively small data set. For performing both, the common factor was bowel preparation, that according to patients was stated the most unpleasant and less favorable due to the constant consumption of fluids and frequent bathroom trips. The majority of people, however, recommended CT since it is less invasive, faster, and easier, and it does not require anesthesia. Later that year, the results of a survey[4] revealed clear evidence of an increase in the incidence of IBD among children in the United States, with more definite evidence of CD than ulcerative colitis. The fact that these findings were discovered had significant ramifications for the disease's diagnosis and treatment. With the wide number of rising cases, the need to find a favorable alternative to colonoscopy and CT exams was looked into. Faecal calprotectin levels were revealed to be a useful screening tool for patients who would benefit from an endoscopy for suspected inflammatory bowel illness in 2010[7]. Adult studies had much stronger discriminative power to safely exclude the disease than studies of children and teenagers. It was discovered that faecal calprotectin levels can provide vital information and assist patient management at the tertiary care level[5]. Machine learning has been extensively used in predicting disease progression over the last decade, including a recent study that identified FC as a significant risk factor for predicting the progression of CD. The best choice of machine learning models is not always obvious, and comparative assessments are required to identify the model with the best predictive performance. The paper published in September 2020[3] is the first of its kind to use FC levels 804 patients to conduct a comparative analysis of the performance of supervised machine learning methods like Logistic Regression, SVM, ANN, Random forest approach for predicting the progression of CD. Based on the model's performance, complexity, and interpretability, it was determined that logistic regression was the best model for predicting Crohn's disease progression in a data set. When compared to the other three machine learning algorithms, logistic regression exhibited a slightly greater accuracy and a significantly higher Area under Curve. It is worth noting that only a small data set was taken under observation in the analysis. On digging deeper into the procedure that these ML algorithms follow, I was able to figure out certain other details regarding the algorithms that I will state ahead. First, logistic regression is useful for detecting SNP-SNP interactions[9] that are linked to the risk of developing a complex disease. The SNP-SNP interactions helps in understanding the genetic origins of complex disease characteristics has long been acknowledged. Identifying SNP-SNP interactions, on the other hand, is computationally difficult. Using parallel computing components, a library can be created to speed up the analysis of SNP-SNP interactions. This library would be an effective instrument for reducing the time it takes to perform logic regression on a computer cluster in applications like SNP analysis and other data analysis. Second, in SVM-based technique[8], to verify the capability of reducing feature space dimensionality, principal component analysis and feature selection techniques are used, and a K-fold cross validation procedure was added into the classifier to better measure results correctness.

Finally, RF categorization reveals that isoleucine and valine levels are higher and lower in CD patients, respectively, than in healthy individuals[2]. Because CD patients are at such a high risk of nutritional deficiency, the alterations in amino acid levels in CD appear to be reasonable. Malnutrition is linked to a decrease in intestinal mucosa function. As a result, determining nutritional status and energy requirements is crucial in the management and follow-up of CD. Further, I was able to look into another research paper that used Machine Learning approaches to predict medication non-adherence in Crohn's disease[10]. In that cross-sectional study, 446 CD patients who had been prescribed AZA were included. The accuracy, recall, precision, F1 score, and area under the curve of two machine learning models, the back propagation neural network (BPNN) and the support vector machine, were constructed and compared with logistic regression. This paper suggested that SVM surpassed back propagation neural networks and logistic regression in practically every aspect. In a nutshell, with the passing of years and noting the constant rise in cases of Crohn's among children as well as adults, the technology to detect the progression of CD became more favorable as it moved from colonoscopies to FC level detection to an advancement like using Machine learning for the same. Moreover, it is pretty evident that as the volume of data set and the parameter change, so does the final output and the efficiency of machine learning algorithms. Despite of this, I am confident that Logistic regression will show very high accuracy as compared to other three algorithms when handling a huge data of patients of Crohn's disease. Along with that, for a comparatively larger data set and more attributes, the accuracy of the other algorithms will also increase greatly.

References

[1] Allison Agus, Sébastien Massier, Arlette Darfeuille-Michaud, Elisabeth Billard, and Nicolas Barnich. Understanding host-adherent-invasiveescherichia coliinteraction in crohn's disease: Opening up new therapeutic strategies. *BioMed Research International*, 2014:1–16, 2014.

[2] Fariba Fathi, Laleh Majari-Kasmaee, Ahmad Mani-Varnosfaderani, Anahita Kyani, Mohammad Rostami-Nejad, Kaveh Sohrabzadeh, Nosratollah Naderi, Mohammad Reza Zali, Mostafa Rezaei-Tavirani, Mohsen Tafazzoli, and et al. 1h nmr based metabolic profiling in crohns disease by random forest methodology. *Magnetic Resonance in Chemistry*, 52(7):370–376, 2014.

[3] Zain U. Hussain, Ragnor Comerford, Fynn Comerford, Nathan Ng, Dominic Ng, Ateeb Khan, Charlie Lees, and Amir Hussain. A comparison of machine learning approaches for predicting the progression of crohn's disease. *2020 IEEE Student Conference on Research and Development (SCOReD)*, 2020.

[4] Janet Z. Liu, Stefan Jellbauer, Adam J. Poe, Vivian Ton, Michele Pesciaroli, Thomas E. Kehl-Fie, Nicole A. Restrepo, Martin P. Hosking, Robert A. Edwards, Andrea Battistoni, and et al. Zinc sequestration by the neutrophil protein calprotectin enhances salmonella growth in the inflamed gut. *Cell Host Microbe*, 11(3):227–239, 2012.

[5] Toshiki G Nakashige, Bo Zhang, Carsten Krebs, and Elizabeth M Nolan. Human calprotectin is an iron-sequestering host-defense protein. *Nature Chemical Biology*, 11(10):765–771, 2015.

[6] Guoqing Ouyang, Guangdong Pan, Qiang Liu, Yongrong Wu, Zhen Liu, Wuchang Lu, Shuai Li, Zheng Zhou, and Yu Wen. The global, regional, and national burden of pancreatitis in 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *BMC Medicine*, 18(1), 2020.

[7] P. F. Van Rheenen, E. Van De Vijver, and V. Fidler. Faecal calprotectin for screening of patients with suspected inflammatory bowel disease: diagnostic meta-analysis. *Bmj*, 341(jul15 1):c3369–c3369, 2010.

[8] Sergio Salerno, Silvia Franchini, Maria Chiara Terranova, and G. Lo Re. Evaluation of a support vector machine based method for crohn's disease classification. *Neural Approaches to Dynamics of Signal Exchanges*, page 314–325, Jan 2020.

[9] Unitsa Sangket, Surakameth Mahasirimongkol, Pichaya Tandayya, Surasak Sangkhathat, Wasun Chantratita, Qi Liu, and Yutaka Yasui. Parallelization of logic regression analysis on snp-snp interactions of a crohns disease dataset model. *Sains Malaysiana*, 46(9):1449–1455, 2017.

[10] Lei Wang, Rong Fan, Chen Zhang, Liwen Hong, Tianyu Zhang, Ying Chen, Kai Liu, Zhengting Wang, and Jie Zhong. Applying machine learning models to predict medication nonadherence in crohn's disease maintenance therapy. *Patient Preference and Adherence*, Volume 14:917–926, 2020.