# 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 13, 2021

**Abstract**

The present research paper is designed to depict a comparative evaluation of performance of four different machine learning algorithms: Logistic Regression, Support Vector Machine, Artificial Neural Network and Random Forest Approach based on their ability to predict the presence of Crohn's disease, provided a large data set of patients with the information regarding their faecal calprotectin level, weight, height, body mass index(BMI) and gender.

**Keywords:** Machine Learning, Logistic Regression, SVM, Faecal calprotectin, Crohn's disease, Parallel Processing

## 1 Introduction

Crohn's disease (CD) is an inflammatory bowel disease that is persistent and gets worse over time. It is distinguished by an uncontrolled inflammation that can affect any part of the gastrointestinal tract, starting from the mouth to anus. Its prevalence has continued to increase over the last 50 years, with northern Europe, the United Kingdom, and North America having the highest occurrence[9]. 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 is the process of breaking down larger problems into smaller, independent, and often similar parts that can be executed concurrently by multiple processors communicating via shared memory, with the results being combined as part of an overall algorithm. The primary focus of parallel computing is to increase the amount of computing power for rapid application processing and problem solving. Parallel computing not only helped enable the study of gut microbiome but also lead to discovery of the bacteria responsible for crohn's disease. Trillions of microbes can be found primarily in the intestine and on skin. The majority of microbes in the intestines are found in a "pocket" of the large intestine called cecum, and they 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 iron, calcium, 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. This research paper provides a comparative analysis of how accurate do four different machine learning methods prove to be, based on their ability to handle a huge data set of patients while predicting the presence of crohn's disease. According to the research and outcome of my work, Logistic Regression proved to be the most accurate and displayed the least possibility of error. The accuracy of the other three algorithms namely Random Forest, Support Vector Machine(SVM) and Artificial Neural Network(ANN) were also evidently good, with Artificial Neural Network displaying a possibility of error higher as compared to the others. First, in Section 2, we will review some previously published work related to the research. We will then address the problem statement and the way by which a solution is provided for it, in Section 3 and Section 4 respectively. Section 5, shows the dissection of the produced output. Lastly, Section 6 states the outcome that was finally generated with the help of our reasoning and mentions some future aspects.

## 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[10] 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[7] revealed clear evidence of an increase in the incidence of irritable bowel disease(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[10]. 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[8].

Machine learning(ML) 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. Most of the times, comparative assessments are required to identify the model with the best predictive performance. The paper published in September 2020[5] 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 was 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[11] 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[4], 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[3]. 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[12]. 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, it can be observed by the end of this research that Logistic regression is exhibiting a higher accuracy as compared to other three algorithms when handling a huge data of patients of crohn's disease. Along with that, for the comparatively large data set used for this research and added attributes, the other algorithms also exhibit a significantly good accuracy.

## 3 Problem Statement

Artificial intelligence (AI) is a new technological advancement in the field of disease prediction and diagnosis that is rapidly gaining traction. In the recent years, AI and machine learning have been shown to help in diagnosing and understanding disease severity, as well as predicting treatment response and disease recurrence. Yet to fully comprehend the working of ML on predicting the outcome of crohn's disease, deep research is required. Lately, large volume of data from electronic health records, clinical trials, medical imaging, and multi-omic databases has been used to improve diagnostic accuracy and treatment response predictability. However, while using big data sets, the high dimensionality of various factors may affect the efficiency of predictive machine learning models. It is rather hard to evaluate the performance of ML methods for certain data sets as sometimes if the data is limited and missing, the accuracy portrayed by these algorithms may not be very efficient as well as precise. This research will help us understand how each machine learning algorithm will perform on a data set of patients, taking into consideration various input factors such as FC level, height, body mass index etc. for predicting CD.

One of the previous work[5] suggested that while predicting the progression of CD over a data set of 804 patients, logistic regression surpassed the others with a high accuracy. The missing data for a number of patients in the data set, lead to inclusion of fewer variables in the model. Moreover, due to the small sample size, SVM, Random Forest, and the ANN did not show superior accuracy in this study when compared to logistic regression. The paper suggested that in future, more variables should be analyzed and the findings should be validated using another large and independent data set. The current research work does exactly that is asked for according to the future aspects of the previous paper. The current research used an enormous data set of 5000 patients having various attributes with no missing values. Along with this, the previous paper noted logistic regression to have the highest accuracy, followed by SVM. Whereas Random Forest and ANN displayed approximately the same efficiency. The result of our research turned out to be slightly different than this. Our output proposed Random Forest to have the second best accuracy after Logistic Regression, besides SVM and ANN following it one after other. On the top of that, this paper also put forward the number of people who tested positive and the number of people who tested negative for crohn's disease out of 5000, with the help of input attributes, which was missing in the previously proposed research.

The future aspect of the 2020 paper[5] necessitated a thorough inspection and validation for which the existing theories and concepts are needed to be studied again to determine the overall scope and betterment of output. There is a requirement to re-examine the performance of ML methodologies for larger volume of data and note how the efficiency and outcome of each method improves. Furthermore, we will see the association of FC level with CD in the upcoming section to know how it helps in predicting the exact number of people who have CD.

## 4 Proposed Solution

In order to conduct a comparative analysis of machine learning algorithms to predict the presence of CD, we first need to know the attributes of the huge data set that are taken for evaluation. The first column contains the FC level, following this are other attributes like weight in terms of kg, height in terms of cm, BMI, gender (female is represented by 1

and male is represented by 0) as input and the last column represents the output of crohn's disease (positive = 1 and negative = 0). The main component here is FC level, so we will understand the relation of it with CD in the forthcoming subsection.

## 4.1 Association of Faecal Calprotectin and Crohn's Disease

Faecal calprotectin is the measurement of the protein calprotectin in the faecal matter of a person. Calprotectin is a kind of protein found in saliva, human blood, urine, and cerebrospinal fluid when inflammation is noticed in some part of the body, but the location of the inflammation cannot always be determined during testing the stated fluids. Calprotectin found in the stool or faeces (also known as FC) has a direct link to bowel mucosal damage that causes intestinal inflammation, such as that caused by an IBD. This activity of increased and abnormal inflammation leads to the elevation of level of faecal calprotectin. Even in an article published in October 2019[6], FC was mentioned as a biomarker of intestinal inflammation. It was discovered that a higher level of faecal calprotectin is linked to a longer-term increase in disease progression, including hospitalisation and surgery. It is worthy to note that this test may be able to replace invasive colonoscopy or radio-labeled white cell scanning in certain clinical scenarios. Coming to the range of FC Level and its association with IBD, another paper[2] performed a detailed analysis on patients of crohn's disease based on their FC level and came to a conclusion that a healthy person's FC level would be of range 50.0 mcg/g or less. When a person's test result shows the level of FC between 50.1 to 250.0 mcg/g, they probably have a borderline inflammatory process going on in their system like a treatable IBD. Moreover, the FC level greater than 250.0 mcg/g was marked as abnormal and suggestive of an active inflammatory process within the gastrointestinal system. This study helped us establish that the patient who would display a value greater than 250.0 mcg/g of FC level, have a high chance of testing positive for CD.

## 4.2 The Machine Learning Algorithms

The four machine learning algorithms used for finding the solution to our problem are Logistic Regression, Support Vector Machine, Random Forest and Artificial Neural Network.

### 4.2.1 Logistic Regression

Logistic Regression is an analytical methodology of machine learning where the final result is a binary variable that is dichotomous in nature. The term dichotomous refers to the fact that there are only two possible classes. In simpler words, it is a classification algorithm that comes into action when the data has binary output, such as when it belongs to one of two classes. It can be also looked at as a variant of linear regression in which the target variable is categorical. The dependent variable is the log of odds and using a logistic function, logistic regression predicts the likelihood of a binary event to occur. The logistic function also referred as the sigmoid function, produces a 'S' shaped curve that is used to map any real-valued number to a value between 0 and 1. For example to predict whether the person has crohn's disease or not or to classify an email as spam or no-spam. The sigmoid function is represented as:

$$SigmoidFunction : Y = \frac{1}{1 + e^{-x}} \tag{1}$$

Here, Y is the output variable, e is euler constant having the value of 2.718 and x is the independent variable. When the sigmoid curve reaches positive infinity, the predicted output becomes 1, and when it reaches negative infinity, predicted output becomes 0. We can classify the result as 1 or YES if the output of the sigmoid function is greater than 0.5, and as 0 or NO if it is less than 0.5. It forms a cutoff line on the plane which divides the data set into two classes as seen in Figure 1, Class A with people who tested positive for CD and Class B with people who tested negative for CD.

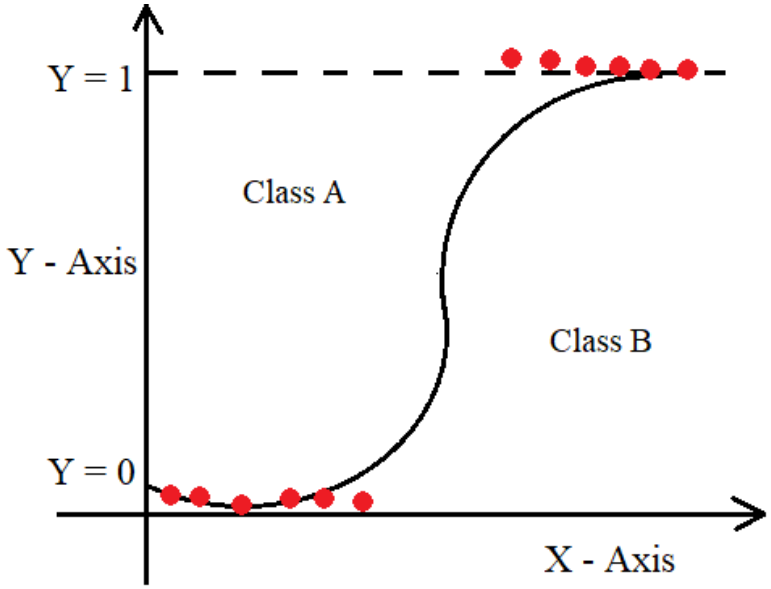


Figure 1: Logistic Regression

#### 4.2.2 Support Vector Machine

SVM is a type of supervised learning method, used for classification and regression analysis. First, we will understand the working of supervised learning with the help of Figure 2. When supervised learning is involved, the input will be in the form of labeled data. This data needs to be analyzed and classified into different classes based on their characteristics, for this, model training is performed on labeled data. If further new data is introduced then it is classified via prediction and the final result will be in the form of two different output classes.

In an effort to divide data set in two classes, SVM has a decision boundary by the name - hyper plane. With all the points scattered across the graph, it is very crucial to notice that a class will have a data point that is very close to the data points of the opposite class. That particular point is termed as the support vector of a class. Similarly, the opposite class will also have a data point that is very close to the data points of the other class and that data point will be the support vector of the opposite class. A line p is constructed close to the support vector of one class and parallel to the hyper plane, similarly another line q is constructed close to the support vector of other class and parallel to the hyper plane. The distance between line p and hyper plane is D- and the distance between line q and hyper plane is D+. Together they form the margin M, therefore M = (D-) + (D+) and margin

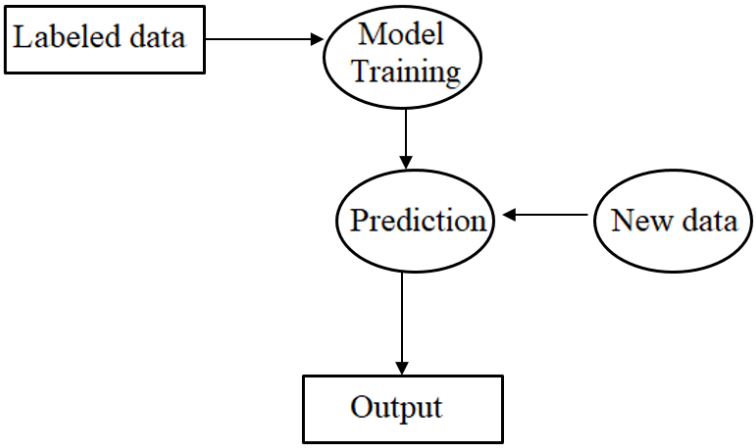


Figure 2: Supervised Learning

decides which hyper plane will exist and which would not.

The above mentioned overall explanation can be simplified by presenting it in the form of an image, as presented in Figure 3.

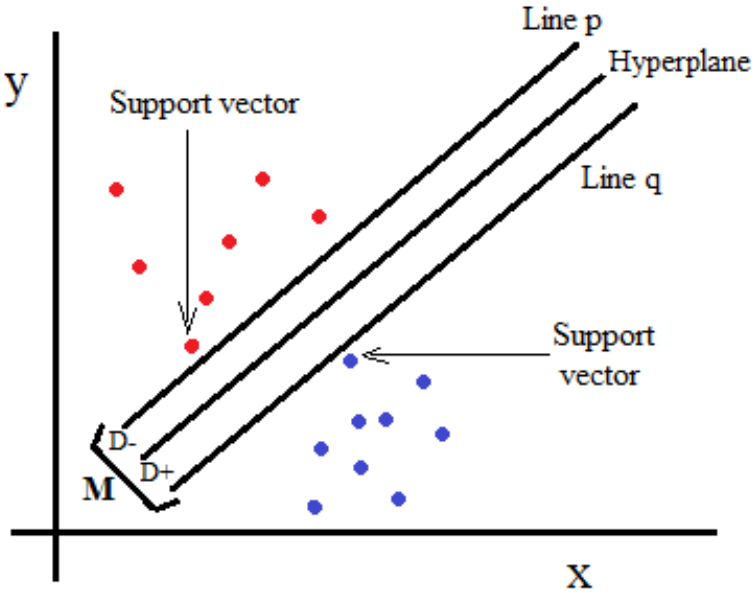


Figure 3: Support Vector Machine

#### 4.2.3 Random Forest Approach

Random Forest is an ensemble classifier that uses decision tree algorithm in a randomized way, where the input is the training data and we have attributes along with one target attribute that has values in the form of Yes and No. A bootstrap data set (BD) is created by randomly picking and inserting any of the samples of training data. In this particular

step, even duplication of data is allowed. Further, a Decision Tree is plotted by selecting nodes in a randomized fashion from the bootstrap data set. Note that, for each selected node, we use subset of variables at each step. An example of Random Forest can be seen in Figure 4: For the root node, the possible variables are A, B, C but for deciding the root node, we consider only 2 variables out of 3 i.e. the subset of total number of variables. Now, by considering two attributes A and B, such that if A is better at splitting the samples, it will be considered as the root node. Subsequently, for the child node, we randomly consider 2 variables, B and C, such that if B is better at splitting the samples then it will be considered as the left child and eventually, C will be the right child, leaving the decision variables, Yes and No to be the leaf nodes. Similarly, for creating more decision trees, the same steps are followed. Finally, we will get a test tuple which decides the value of target attribute with the help of final values of attributes A, B and C. This tuple is applied to every decision tree to know the opinion of each one of them. Based on the majority opinion of each decision tree, we will get the value of target attribute of the test tuple. To summarize, the RF models combine the results of a series of randomly created regression decision trees to predict the final output. They are considered to be reliable predictors for data with high dimensionality as well as small sample sizes.

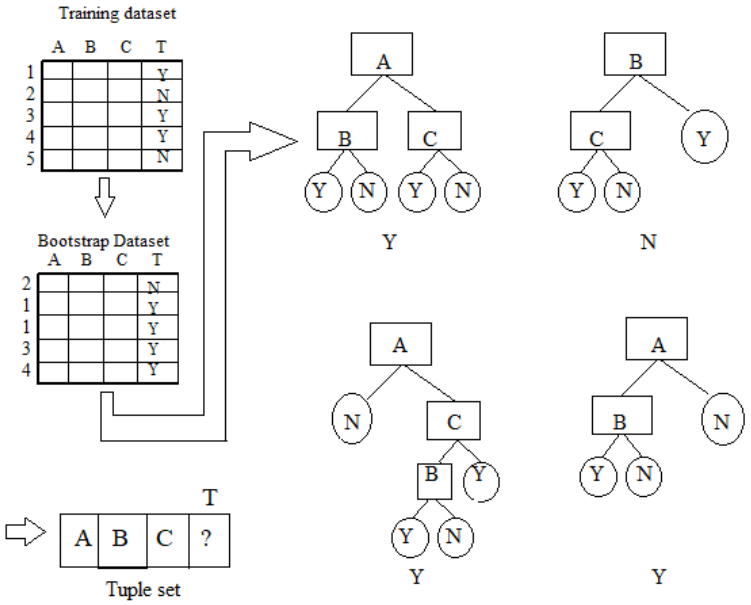


Figure 4: The working of Random Forest approach

### 4.2.4 Artificial Neural Network

An artificial neural network is a connection of neurons where each neuron in the network receives signals from various other neurons, this means that a particular neuron basically receives its input which can originally be the output of some other neuron. Let us understand Figure 5 that considers x1, x2...xn as input signals coming from various other neurons. Each of these signals have an associated weight, denoted by w1, w2...wn. The neuron is divided into two parts: the summation ($\sum$) and the activation function (f), here $\sum$ evaluates (x1w1 + x2w2... xnwn), whereas an activation function calculates a weighted sum and then adds

bias to it to determine whether a neuron should be activated or not. There are 5 kinds of activation functions: Linear, Sigmoid, Tanh, RELU and Softmax. ANN can be used to streamline the diagnostic process in daily life and avoid misdiagnosis. This adaptive learning algorithm can deal with a wide range of medical data and combine it into classifiable outputs.

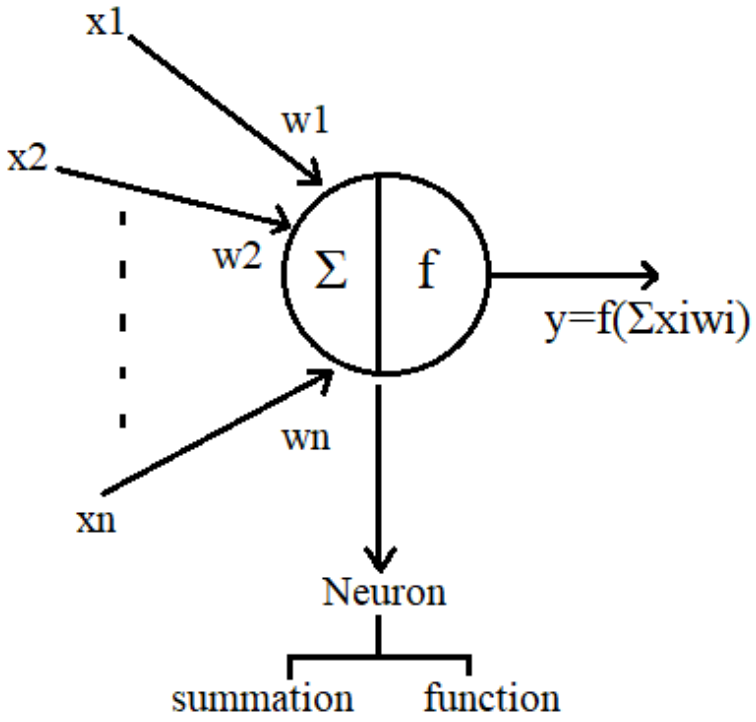


Figure 5: A demonstration of Artificial Neural Network

### 4.3 Dissection of the Code

As mentioned before, this research used a six columned data set for handling the information of 5000 patients related to their FC Levels, weight, height, BMI, gender and result of crohn's. We will now look at the programming done on this data to produce the required output.

#### 4.3.1 The Programming

Our research focuses on machine learning algorithms, hence Python is used as the programming language, while the environment used for running our code is Google Colaboratory also known as Google Colab. Google Colab is a web-based environment that allows us to write and run arbitrary Python code. It is essentially a hosted Jupyter notebook service that does not require any kind of setup and provides a free access to computing resources like GPUs. The libraries utilized while coding are as below:

1. NumPy - It is a third-party library that provides support for huge, N-dimensional array object to store data. It also provides multiple mathematical functions to operate those arrays of numbers.

2. multiprocessing – Multiprocessing refers to the to a machine's ability to run multiple processors at the same time. This library helps to take full advantage of multiple processors on a single machine.

3. pandas - The primary function here is data analysis. Pandas supports the import of data from a variety of file formats for data analysis and manipulation.

4. matplotlib.pyplot – It is a Python library used for plotting 2D graphics.

5. sklearn library that has inbuilt classifiers for LR, SVM, ANN and Random Forest like LogisticRegression, SVC, MLPClassifier, RandomForestClassifier respectively.

As seen in Figure 6, in the first line of code, we have loaded our "Crohn's Dataset.csv" file, and as it is in numeric format, we have further specified what each column represents. Ahead of this, we have used dataframe - the data structure of NumPy library and made it read the tabular data present in the "Crohn's Dataset.csv" file. In addition, we have given the values stored in dataframe to an array. We have initialized two arrays: X and Y. X stores the data values of the whole data set except the last column, which is on the other hand stored in Y i.e. Y stores the last column holding the results of patients in the form of 0 and 1, with 0 representing Negative for CD and 1 representing Positive for CD. In the rest of the code, we have also defined fcl that stores the values of first column of data set representing FC Level. FC level has a major impact on detecting CD progression, it is one of the primary attribute of the input fields due to its ability to recognize the presence of the disease. The number of patients out of 5000 who have tested positive as well as tested negative for CD are also mentioned in the output section, which was done with the help of values of FC level and its association factor with CD.

```
url = "Crohn's Dataset.csv"
names = ['FC Level', 'weight', 'height', 'BMI', 'gender', 'crohns']
dataframe = pandas.read_csv(url, names=names)
array = dataframe.values
X = array[:,0:5]
Y = array[:,5]
```

Figure 6: A part of the code

In the next part of my code, each ML model is prepared and evaluated. Later, k-fold validation is performed with n splits. K-fold randomly divides the data set into n batches. Out of these n batches, some rows are randomly selected as training data and the rest are selected as testing data. After all this evaluation, we will get the mean accuracy and the standard deviation accuracy for each of the four machine learning algorithms.

Finally, there are two types of execution in parallel processing: synchronous and asynchronous. For our code, we have used synchronous execution that occurs when all processes are completed in the same order they were started, this can be achieved by locking the main program until the processes in the question are completed. To execute parallel processing, we have used the Pool class in the multiprocessing library. Pool() works in the background, running multiple Python processes and distributing your computations across multiple CPU cores, allowing them to run in parallel. It provides the Pool.apply() method, which accepts the parameters passed to the 'function-to-be-parallelized' as an argument in the args argument.

# 5 Experimental Evaluation

The final output of the whole above programming is represented in the form of a bar graph that portrays the mean accuracies of logistic regression, support vector machine, random forest and artificial neural network in the same order as mentioned here. Although it is clear that the performance of the logistic regression is finest, there is a considerable low difference in the performance of each algorithm. The mean accuracy for this research represents the amount of true negative and true positive that an algorithm is likely to predict and the standard deviation accuracy shows the amount of possible error in the prediction also stated as the false negative and false positive outcomes. If an algorithm have high mean accuracy, it will produce more volume of true negative and true positive data but if an algorithm has high standard deviation accuracy, it will produce more volume of false negative and false positive data.

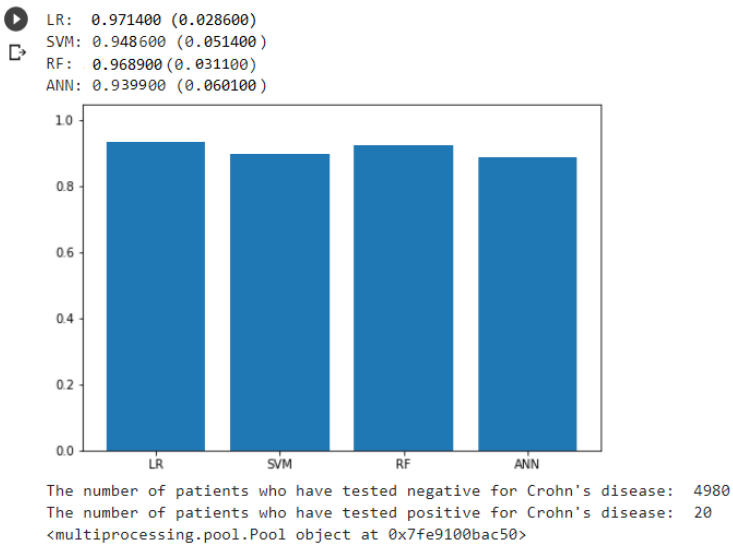


Figure 7: Output in the form of graphical representation for Mean Accuracy

On the other hand, Figure 8 is a depiction of the comparative analysis of standard deviation accuracies of ML algorithms. It is plausible to say that ANN has the SD accuracy of 6.01 percent which means it has the highest chance of producing an error, following it is SVM that illustrates the possibility of 5.14 percent of error. Random forest methodology with a 3.11 percent of SD accuracy and logistic regression with a surprisingly low SD accuracy of 2.86 percent can be termed relatively efficient and better at handling huge medical data sets.

In a nutshell, the algorithm that represents the highest mean accuracy and least standard deviation accuracy would be most accurate. On the contrary, the algorithm that represents the least mean accuracy and highest standard deviation accuracy would be the least accurate. The comparative difference between both accuracy of every single algorithm is illustrated in Figure 9.

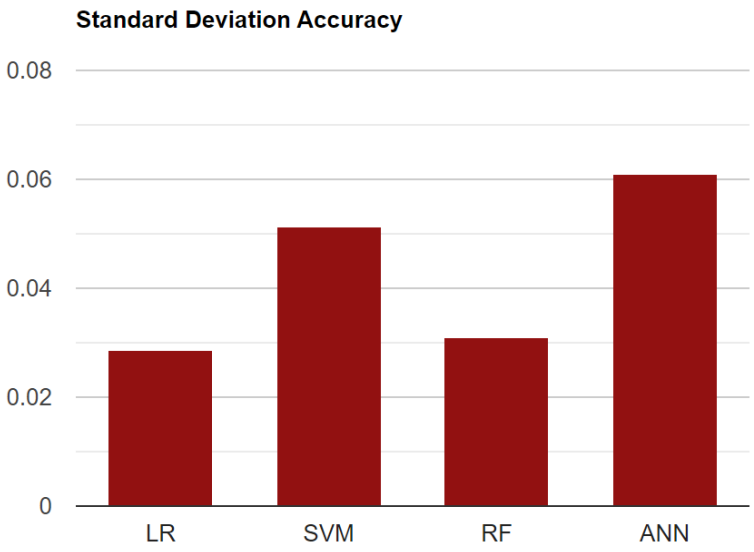


Figure 8: The graphical representation for Standard deviation Accuracy

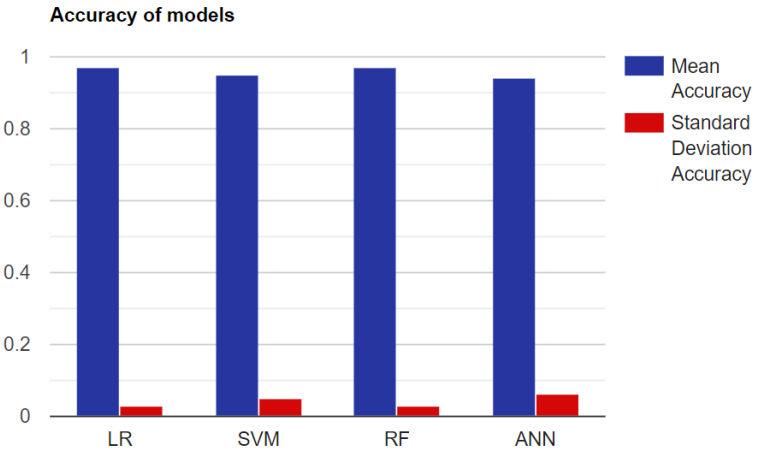


Figure 9: Accuracy of models

## 6 Conclusions

As a consequence of consistently escalating cases of crohn's disease, machine learning algorithms were put into testing and evaluation in the recent years which lead to various experiments on these algorithms in the domain of disease diagnosis. In this research, different machine learning algorithms have been studied and assessed , in order to differentiate between their potential to perform efficiently on a large volume of data.

Logistic Regression predicted the highest accuracy for a large data set of patients, followed by Random Forest with the accuracy of 96.89 percent. Support Vector Machine also displayed a considerably good accuracy, whereas for a huge data set as the one used for this research, Artificial Neural Network portrayed the least accuracy as compared to the other models.

The scope of disease diagnosis is constantly increasing with the aid of machine learning algorithms. Although the performance and accuracy of every individual algorithm is quite high, but there is a likelihood that two different algorithms combined and made to handle a large data set would produce an even superior result. Furthermore, it is crucial to take into consideration that with every country and region, the resident's food habits and diet changes. Due to the pattern of their eating habits, the collected data having the sample values and results would almost be constant and there would be very less variety. In this case, if the data of groups of people from various parts of the world is collected and merged together to form an enormous data set, the outcome and accuracy might end up being even more appropriate.

# 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] A Dhaliwal, Z Zeino, C Tomkins, M Cheung, C Nwokolo, S Smith, C Harmston, and R P Arasaradnam. Utility of faecal calprotectin in inflammatory bowel disease (ibd): what cut-offs should we apply? *Frontline Gastroenterology*, 6(1):14–19, 2014.

[3] 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.

[4] Silvia Franchini, Maria Chiara Terranova, G. Lo Re, and Sergio Salerno. Evaluation of a support vector machine based method for crohn's disease classification. *Smart Innovation, Systems and Technologies*, page 313–325, 2020.

[5] 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.

[6] Nicholas A. Kennedy, Gareth-Rhys Jones, Nikolas Plevris, Rebecca Patenden, Ian D. Arnott, and Charlie W. Lees. Association between level of fecal calprotectin and progression of crohns disease. *Clinical Gastroenterology and Hepatology*, 17(11), 2019.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.