Predicting IBD risk from genomic data using machine learning

Samit Watve, BrainStation Capstone Project

Introduction:

Inflammatory bowel disease (IBD) is a debilitating disorder which causes severe pain, inflammation, and bleeding in the digestive system. IBD affects 1.6 million people in North America and leads to $31B in annual healthcare costs in the US [1]. IBD is manifested in two main varieties: Crohn's Disease (CD) which can affect the entire digestive system and Ulcerative Colitis (UC) which is typically limited only to the large intestine (colon) [2]. The primary risk factors for developing IBD include diet, history of smoking, microbiome and genetics [3]. The motivation behind this project was to develop a predictive algorithm capable of distinguishing healthy and affected individuals based solely on genetic data. The primary advantages of using genetic data include for this purpose include: 1) Stability of genomic information, since the genome is fixed at conception and does not change through an individual's lifetime; 2) Availability of large, annotated genomic datasets; 3) Steadily decreasing costs of obtaining genomic information at scale. Prior studies have looked at single nucleotide polymorphism (SNP) data across the genome to look for associations between specific genes and disease risk [4], [5], [6]. One well known example of a gene associated with IBD is NOD2 which is intestinal receptor that recognizes peptidoglycans on bacterial cell surfaces and is involved in initiating inflammation when exposed to bacterial cells [7].

Machine learning approaches have been used previously to distinguish between CD/ UC patients using SNP data with decent success [8], [9]. However, these studies used proprietary data which is not open source and therefore difficult to replicate. In this project I used several binary classification algorithms for predicting whether a given individual belongs to the healthy or the diseased (IBD) class using select SNP data as input features. Logistic regression and Multi-layered Perceptron models performed the best with accuracy scores ranging between ~0.81-0.84 across several trials and ROC-AUC values of ~0.9 or greater, which is higher than some previously published models [8], [9]. This project demonstrates the viability of machine learning approaches for disease risk classification and adds another tool to the clinician's toolbox. Such methods will enable researchers to develop personalized interventions at an early age before disease symptoms appear.

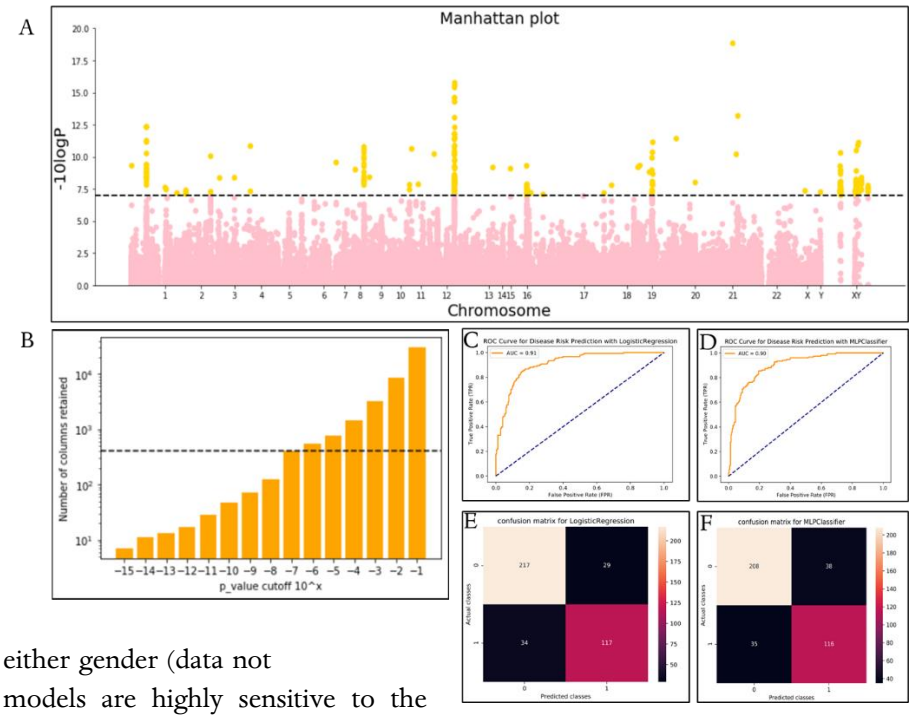
Methods:

Data files were sourced from https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/11991, which were generated for a scientific study on early onset IBD in the European population [10].This dataset contained demographic and genetic information for 2645 patients which included DNA sequence information on ~200,000 single nucleotide polymorphisms (SNPs). All code was written in Python v3.7.8 using a Jupyter lab (2.2.9) environment. Data cleaning, pre-processing and visualization was performed using data science packages such as pandas (1.1.4), numpy (1.19.4), matplotlib (3.3.3) and seaborn (0.11.0), and machine learning was performed using sci-kitlearn (0.23.2) or xgboost(1.3.0.post0). After initial EDA and data cleaning, SNP filtering was performed using a custom data cleaning script that performed over 100,000 chi-squared tests using the scipy.stats (1.4.1) module. Only the top features that had a p-value (< 1e-7) were retained for further analysis. All features in the dataset such as gender, the target healthy / disease class and the retained SNP features were One-Hot encoded thereby eliminating the need for feature scaling. Since p-value filtering already reduced the dataset down to a small number of features, further dimensionality reduction was not implemented. Several binary classification algorithms such as logistic regression, multi-layer perceptron,

xgboost, KNN etc were trained using 85% of the data and scored for their ability to correctly predict healthy or diseased individuals from the test set (15% of the data). Hyperparameter tuning was performed using gridsearchCV using 5-fold cross-validation. Models were also evaluated using precision, recall and ROC-AUC metrics.

Results and discussion:

After implementing the data cleaning and processing pipeline, the number of SNP features was reduced from 196523 to just 413 features that had statistically significant associations between a given SNP and the healthy or the diseased class (fig 1A; gold dots). The features that were retained were distributed across the entire genome, however a few hotspots were observed on chromosomes 1, 8, 12, 19 and Y respectively. Given the aggregation of features on the Y chromosome, I postulated that gender might play a significant role in determining disease risk. However, there was no statistically significant difference observed in disease predilection for either gender (data not shown). Machine learning models are highly sensitive to the number of input features that are used for model training. Therefore, the number of features to be included, is a tunable parameter in the context of machine learning. For this dataset there is a direct relationship between the number of features retained and the p-value cutoff (fig 1B). By using a stricter cut-off, we can further restrict the number of features that go into the model fitting pipeline. In my experiments, the use of a less stringent p-value cutoff (i.e., a higher number of features) did not significantly improve model accuracy (data not shown). Indeed, using a high number of features led to overfitting on the training data, as well as poor accuracy scores on the validation data. Therefore, only these 413 features were used for further optimization steps. This process involved iterating over a range of hyperparameters pertaining to several binary classification algorithms. Models that obtained the highest prediction accuracy scores after 5-fold cross validation were retained for further analysis. Across multiple trials, Logistic regression ('C': 1.0, 'penalty': 'l2', 'solver': 'lbfgs') and Multi-layered Perceptron ('alpha': 0.1, 'hidden_layer_sizes': 10, 'max_iter': 1250, 'solver': 'sgd') models consistently performed the best with accuracy scores ranging between ~0.81-0.84 (see table 1 below for a full summary). The preferred metric for evaluating model performance in the context of disease risk prediction is the ROC-AUC. A perfect diagnostic would have an AUC of 1.0, however in practice values close to 1.0 are difficult to achieve. A previous study applied Logistic regression with l1 penalty on a dataset (that included a total of 60,828 samples) to distinguish between healthy / UC or healthy / CD patients reported AUC values of 0.83 and 0.86 respectively [8]. Similarly, another publication compared the performance of various LR models such as Lasso, Ridge and



ElasticNet as well as GBT methods like XGBoost, LightGBM and CatBoost and reported a maximum AUC of 0.80 using similar sample sizes [9]. In this study, Logistic regression ('C': 1.0, 'penalty': 'l2', 'solver': 'lbfgs') had the highest AUC of 0.91, followed by MLP with AUC values that ranged between 0.85-0.90 across multiple trials. Additionally, neither model skewed heavily toward classifying individuals as healthy or diseased, despite the class imbalance (~1600 healthy vs ~1000 diseased) within the dataset (see confusion matrixes in fig 1E and F) and had high precision and recall values for either class (Table 1). This suggests that both models "learned" to correctly distinguish between the classes as opposed to randomly guessing one class for a high proportion of the cases. While this performance is impressive, I was working with smaller sample sizes compared to the studies listed above. Additionally, the studies referenced above included >10,000 features for their final model fitting. Thus, direct comparisons between the different studies may not be meaningful. Nonetheless, I was able to achieve relatively strong model performance using an ordinary laptop in a relatively short amount of time. In future projects, I will try to explore the effect of including feature data from a variety of sources such as diet, microbiome, and family history. The machine learning approaches used here can also be extended to develop disease risk prediction algorithms for disorders other than IBD.

Model	Accuracy	Precision (healthy / IBD)	Recall (healthy / IBD)	ROC-AUC	Best model Parameters
Logistic Regression	0.84	0.86 / 0.80	0.88 / 0.79	0.91	'C': 1.0, 'penalty': 'l2', 'solver': 'lbfgs'
Multi-layer perceptron	0.82	0.86 / 0.75	0.85 / 0.77	0.90	'alpha': 0.1, 'hidden_layer_sizes': 10, 'max_iter': 1250, 'solver': 'sgd'
Xgboost	0.79	0.82 / 0.73	0.85 / 0.65	0.88	'learning_rate': 0.1, 'n_estimators': 30
Gradient Boosting	0.79	0.80 / 0.74	0.86 /0.66	0.87	'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 35
SVC	0.80	0.82 / 0.75	0.85/ 0.70	0.84	'C': 1.0, 'gamma': 'scale', 'kernel': 'poly'
Ridge Classifier	0.82	0.85 / 0.77	0.86 / 0.75	0.76	'alpha': 0.9
Random Forest	0.70	0.73 / 0.63	0.82 / 0.51	0.76	max_features': 'sqrt', 'n_estimators': 1000
K-Nearest Neighbors	0.60	0.64 / 0.46	0.81 / 0.26	0.61	metric': 'manhattan', 'n_neighbors': 11, 'weights': 'distance'

Table 1- Model performance summary statistics for various binary classifiers

References:

1. https://www.mayoclinic.org/diseases-conditions/inflammatory-bowel-disease/symptoms-causes/syc-20353315
2. https://www.cdc.gov/ibd/data-statistics.htm#:~:text=Inflammatory%20Bowel%20Disease%20Prevalence%20(IBD,%25%20or%202%20million%20adults).
3. McDowell C, Farooq U, Haseeb M. Inflammatory Bowel Disease. [Updated 2020 Jun 28]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470312/
4. Ek, W. E., D'Amato, M., & Halfvarson, J. (2014). The history of genetics in inflammatory bowel disease. Annals of gastroenterology, 27(4), 294–303.
5. Chen, G. B., Lee, S. H., Montgomery, G. W., Wray, N. R., Visscher, P. M., Gearry, R. B., Lawrance, I. C., Andrews, J. M., Bampton, P., Mahy, G., Bell, S., Walsh, A., Connor, S., Sparrow, M., Bowdler, L. M., Simms, L. A., Krishnaprasad, K., International IBD Genetics Consortium, Radford-Smith, G. L., & Moser, G. (2017). Performance of risk prediction for inflammatory bowel disease based on genotyping platform and genomic risk score method. BMC medical genetics, 18(1), 94. https://doi.org/10.1186/s12881-017-0451-2

6. Xue, A., Wu, Y., Zhu, Z. et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. Nat Commun 9, 2941 (2018). https://doi.org/10.1038/s41467-018-04951-w
7. Negroni, A., Pierdomenico, M., Cucchiara, S., & Stronati, L. (2018). NOD2 and inflammation: current insights. Journal of inflammation research, 11, 49–60. https://doi.org/10.2147/JIR.S137606
8. Large Sample Size, Wide Variant Spectrum, and Advanced Machine-Learning Technique Boost Risk Prediction for Inflammatory Bowel Disease Wei, Zhi et al. The American Journal of Human Genetics, Volume 92, Issue 6, 1008 – 1012 https://www.cell.com/ajhg/fulltext/S0002-9297(13)00215-2#secsectitle0030
9. Romagnoni, A., Jégou, S., Van Steen, K. et al. Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. Sci Rep 9, 10351 (2019). https://doi.org/10.1038/s41598-019-46649-z
10. Dissecting Allele Architecture of Early Onset IBD Using High-Density Genotyping Cutler DJ, Zwick ME, Okou DT, Prahalad S, Walters T, et al. (2015) PLOS ONE 10(6): e0128074. https://doi.org/10.1371/journal.pone.0128074