**Supervised machine learning classifies inflammatory bowel disease patients by subtype using whole exome sequencing data**

Imogen S Stafford1,2,3\*, James J Ashton1,4\*, Enrico Mossotto1, Guo Cheng1,2, R Mark Beattie4, Sarah Ennis1

1. Department of Human Genetics and Genomic Medicine, University of Southampton, Southampton, UK
2. NIHR Southampton Biomedical Research, University Hospital Southampton, Southampton, UK
3. Institute for Life Sciences, University of Southampton, Southampton, UK
4. Department of Paediatric Gastroenterology, Southampton Children’s Hospital, Southampton, UK

\*denotes joint 1st authors

Correspondence to-

Professor Sarah Ennis

Human Genetics & Genomic Medicine

University of Southampton

Duthie Building (Mailpoint 808)

Southampton General Hospital

Southampton

SO16 6YD

Tel: +44 (0)23 8079 8614

Fax: +44 (0)23 8079 4264

[s.ennis@southampton.ac.uk](mailto:s.ennis@southampton.ac.uk)

Word Count: 3,146

Abstract Word Count: 248

**ABSTRACT**

**Background**: Inflammatory bowel disease (IBD) is a chronic inflammatory disorder with two main subtypes: Crohn’s disease (CD) and Ulcerative Colitis (UC). Prompt subtype diagnosis enables the correct treatment to be administered. Using genomic data, we aimed to assess machine learning (ML) to classify patients according to IBD subtype.

**Methods**: Whole exome sequencing from paediatric/adult IBD patients was processed using an in-house bioinformatics pipeline. This data was condensed into the *per-gene, per-individual* genomic burden score, GenePy. Data was split into training and testing datasets (80/20). Feature selection with a linear support vector classifier, and hyperparameter tuning with Bayesian Optimisation was performed (training data). The supervised ML method random forest was utilised to classify patients as CD or UC using three panels: I) all available genes, 2) autoimmune genes, 3) ‘IBD’ genes. ML results were assessed using AUROC, sensitivity and specificity on the testing dataset.

**Results**: 906 patients were included in analysis (600 CD, 306 UC). Training data included 488 patients, balanced according to the minority class of UC. The autoimmune gene panel generated the best performing ML model (AUROC=0.68), outperforming an IBD gene panel (AUROC =0.61). *NOD2* was the top gene for discriminating CD and UC, regardless of the gene panel used. Lack of variation in genes with high GenePy scores in CD patients was the best classifier of a diagnosis of UC.

**Discussion:** We demonstrate promising classification of patients by subtype utilising random forest and WES data. Focussing on specific subgroups of patients, with larger datasets may result in better classification.

Keywords: inflammatory bowel disease, machine learning, genomics

**INTRODUCTION**

Inflammatory bowel disease (IBD) is a complex, heterogenous, immune-mediated condition, which may be considered of autoimmune genetic aetiology. It is characterised by chronic relapsing and remitting inflammation in the gastrointestinal tract. Crohn’s disease (CD) and ulcerative colitis (UC) are the two major diagnostic subtypes, discriminated largely by disease location and histological findings. Overlapping features, such as isolated colonic inflammation, can impair discrimination of subtypes and some patients remain unclassified (Inflammatory Bowel Disease Unclassified – IBDU).1 Delayed subtype diagnosis can result in an increased risk of complications that can require surgery.2,3 Current practice requires endoscopic, histological and radiological assessment, alongside clinical judgement, to differentiate between Crohn’s disease and ulcerative colitis4. Prompt and accurate diagnosis can be particularly important for paediatric patients: a delay of over 8 months has been shown to be independently associated impaired growth that persists one year after diagnosis.5

IBD is considered a complex genetic condition, with disease susceptibility derived from a combination of multiple genes interacting with the environment. Linkage and association studies at the end of the last century identified *NOD2* variation as the biggest risk factor for developing CD.6 IBD was a prominent focus of many genome-wide association studies (GWAS) and successfully led to the identification of more than 200 genes impacting risk.6,8 These genes were enriched in molecular pathways with a role in innate and adaptive immunity, autophagy, *IL10* signalling and epithelial barrier integrity.9 These data have improved insight into the molecular aetiology of disease at a population level, although there were comparatively few loci associated with a specific IBD subtype.

High throughput genomic sequencing is rapidly transforming precision diagnostics to inform targeted medicine in both rare disease and cancer. National programmes such as the Genomics England 100,000 Genomes Project and the All of Us project within the US-based Precision Medicine Initiative, are driving this technology into mainstream medical practice. 10,11 To date there has been less focus on the clinical application of either whole genome sequencing (WGS) or whole exome sequencing (WES) in diseases perceived to be genetically complex. The advent of the National Health Service’s Genomic Medicine Service (GMS) demonstrates that in the near future it is possible for sequencing to become routine for patient assessment. Already, patients with early-onset IBD can be referred to the GMS for sequencing, as these individuals are more likely to be diagnosed with a primary immunodeficiency with an IBD-like phenotype. However, for other patients with an oligogenic or polygenic disease aetiology, interpretation of these complex disease genomics is hugely challenging.

Machine learning is a contemporary branch of statistics suited for analysis of high-dimensional biological data. Supervised machine learning (ML) algorithms discover patterns in data variables that are associated with specific outcome labels. These learned patterns can then be applied to new data, and the algorithm then predicts the outcome without knowledge of the label. Complex biological data is known to have intrinsic technical and biological noise that can reduce the performance of most classification algorithms. 12,13 Dealing with dimensions (variables) much larger than the number of samples is a challenge in machine learning approaches that can be addressed through dimensionality reduction (*e.g*. principal component analysis), data regularisation (*e.g.* LASSO regularization) or feature selection methods such as prior knowledge application and recursive feature elimination.

We have noted before that ML has been applied to many autoimmune diseases, in order to develop algorithms that can stratify patients. 14 In fact, its application to IBD has only increased in recent years.15 Previous ML models have revealed the potential to classify patients according to their IBD subtype using clinical data.16 However, the potential of genomic data to classify patients by subtype remains understudied.15 We utilise WES data for classification of a cohort of paediatric and adult IBD patients into their disease subtypes CD and UC using a random forest ML algorithm. We employ the pathogenicity burden score algorithm GenePy to transform this large-scale, complex genomic variant data into a single gene score utilising zygosity, allele frequency, and predicted pathogenicity. Larger GenePy scores therefore reflect a higher burden of rarer, deleterious variants.

**METHODS**

Sample data

Inflammatory bowel disease patients were recruited through the Southampton Genetics of IBD study at Southampton Children’s Hospital and University Hospital Southampton. Paediatric patients were diagnosed according to the modified Porto criteria17, and adult patients diagnosed according to British Society of Gastroenterology guidelines. 18 Genomic DNA was extracted from peripheral blood and collected in EDTA by the salting out method. DNA was fragmented, and enriched with Agilent SureSelect All Exon capture kit (version 5 or 6). Libraries were then sequenced on Illumina HiSeq systems. At the time of analysis, no patients with a confirmed diagnosis of a monogenic form of inflammatory bowel disease were included.

Clinical data

All patient diagnoses were reviewed through the electronic health record prior to inclusion, so that the most up to date diagnosis was included subsequent models. Uncertain diagnoses, not fulfilling criteria for CD or UC were termed IBDU and not included in the analysis.

Ethical approval

The study has ethical approval from Southampton & South West Hampshire Research Ethics Committee (09/H0504/125).

Sequencing data processing

Raw whole exome sequencing data were aligned against the human reference genome (GRCh38) with HLA decoys using BWA-mem aligner (v.0.7.17) and duplicate reads were marked for each individual sample19. Samples were individually called with GATK’s (v.4.1.2) 20 HaplotypeCaller and GenotypeGVCFs for the genomic region defined by the union of the two capture kit, with 150 base pair padding. Then all samples were joint called using GATK’s GenomicsDB and GenotypGVCFs to generate a cohort variant call format (VCF) file. Variant Quality Score Recalibration (GATK v.4.1.2) tranche thresholds were identified in the cohort VCF file for single nucleotide variants and indels separately (see https://github.com/UoS-HGIG/WES\_multicalling\_pipeline\_2020).

Prior to annotation, the joint call VCF file was restricted to the intersection of version 5 and 6 capture kits to harmonise the data. The cohort VCF file was then annotated with the gnomAD v.2.1.1 allele frequency across all populations, deleteriousness metric CADD (V.1.6) 21, and a gnomAD flag that indicates technical noise, using Ensembl-VEP (v.103). The annotated VCF was filtered to ensure high quality data, using the approach described by Carson et al22. Individual calls with a genotype quality (confidence) < 20, and a variant depth < 8 were filtered out using VCFtools v.0.1.16. Additionally, variants with a mean genotype quality across the cohort < 35, a call rate lower than 88%, or were recorded as technical noise by the gnomAD flag23 were excluded. We then retained all variant sites with only one alternate allele in the cohort that passed the 0.99 tranche of the VQSR.

We then transformed this variant data into gene-level, per-sample scores using the GenePy scoring system previously described24. In summary, variants are weighted and incorporated according to their frequency in the general population (gnomAD), their observed zygosity, and predicted deleteriousness (CADD). We included exonic variants with high predicted deleteriousness scores (CADD Phred ≥ 15) in the generated GenePy matrix. Further information can be found at <https://github.com/UoS-HGIG/GenePy-1.4>.

Supervised machine learning

The ML pipeline was constructed in Python (v.3.7). Before ML, the GenePy matrix and patient data were filtered to ensure high quality, unbiased data was used as ML input. Genes with no variation were removed, as well as genes that were present on a remapped list of genes identified as false positives in diagnostic genomics25 (Supplementary File 1). Patient ancestry was predicted using Peddy26, and modelling was performed on patients with a high confidence prediction (probability > 0.9) of European ancestry. Related patients were also identified and for each pair of related individuals, we retained the younger patient at diagnosis for downstream analysis.

A random forest (RF) classifier performed supervised ML modelling to classify patients as UC or CD. RF models previously demonstrated superior performance in modelling complex biological data, where usually the number of features greatly exceeds the number of samples.27 These performances are mostly attributable to the intrinsic cross-validation logic on which RF models are based. 28 RF modelling was performed utilising three gene panels: 1) all genes we could generate GenePy scores for, 2) a commercial autoimmune gene panel curated by HTGEdgeSeq, this autoimmune panel of genes was independently compiled to include genes involved in type I and II interferon response, innate and adaptive immune-related interleukins, tumour necrosis factor pathways, toll-like receptors, immune cell signaling, immune checkpoint and co-stimulatory targets and additional immunomodulatory agents, by HTG molecular. This panel has been previously utilised and validated in inflammatory bowel disease genetic research to reduce the number of genes inputted into models to improve biological insights and model performance29,30, 3) an in-house curated IBD gene panel comprised of genes identified in IBD GWAS combined with genes implicated in monogenic IBD (Supplementary File 1). To develop the IBD gene panel we reviewed 298 IBD GWAS loci through database review of the GWAS catalogue (https://www.ebi.ac.uk/gwas/). Utilising these data we mapped the loci to all potential IBD-association candidate genes through genetic mapping and impact on transcription. We combined this list with all genes implicated in monogenic forms of IBD to give a list of protein-coding genes implicated in IBD31. The pipeline was implemented in Python (v.3.7), using scikit-learn.32

Input data were split into training and test datasets. The training set consisted of 80% of the minority class data, and a matching number of samples in the majority class. The remaining data became the test dataset. . This method was used due to awareness of the sensitivity of machine learning models to imbalanced classes (over representation of one class) in datasets.. Then, GenePy scores in the training data were scaled by the maximum score of each gene (MaxAbsScaler), and this scaling model applied to the test data. GenePy scores have variable scoring scales depending on the mutational burden in each gene, hence normalisation between 0 and 1.

Feature selection to discover the optimal set of genes for the classification task was performed with the training set on each of the 3 gene panels. We used a linear support vector classifier (LinearSVC) with a regularisation parameter (C) of 1 and an L1 regularisation penalisation within a 10-fold cross validation (CV) scheme. Genes associated with a coefficient of zero in all 10 folds of the LinearSVC were removed from the training data. Using this training data with the genes chosen by feature selection, hyperparameter tuning was performed using Bayesian optimisation. In a nested CV scheme (7-fold outer CV, 5-fold inner CV), BayesSearchCV chose hyperparameters values that optimised the Random Forest algorithm for this dataset. The optimal hyperparameter value combination was chosen according to the highest, consistent balanced accuracy for the inner and outer folds.

Finally, the random forest classifier was trained with the optimal hyperparameter values, using the genes chosen by feature selection, with the training dataset. The random forest model was then applied to the test dataset. ML model performance was assessed by observing its performance on the test dataset using several metrics: precision, sensitivity, specificity, F1 statistic and Area Under the Receiver Operating Characteristics curve (AUROC). Features were ranked by relative importance to classification, allowing the most discriminating genes to be determined. We utilised SHAP (SHapley Additive exPLanations) values33, to explain the contribution of individual genes to the ML model classification. This ML process was repeated for each of the three gene panels. The full ML pipeline is illustrated in Figure 1, with the coding script available in Supplementary File 2.

*NOD2* to differentiate CD *vs* UC

We hypothesised that *NOD2* would be the main discriminatory factor between CD and UC. We tested the ability of the *NOD2* GenePy score alone to differentiate between patients with CD and UC by using the training and testing sets. We performed iterative Fisher’s exact tests in the training data to determine the optimal *NOD2* GenePy score value to differentiate between CD and UC, and also performed an AUROC analyses to determine the accuracy, sensitivity and specificity of this value. We applied this cut-off to the testing data and determined the ability to differentiate between CD and UC using a *ꭓ2* test.

**RESULTS**

Patient and genomic data characteristics

The cohort included 1,079 individuals diagnosed with IBD, of which 577 were diagnosed at under 18 years, and 495 were diagnosed as adults. Table 1 details the total number of patients in the training and testing dataset, after pre-processing steps (European ancestry only, exclude related patients, only CD or UC). The GenePy matrix was constructed from 135,867 exonic variants, and GenePy scores were generated for 15,669 genes. After pre-processing the genomic data (scores with variance, exclude false positive genes), the number of genes used as input for machine learning is as follows:

1. All available genes: 14,922
2. Autoimmune gene panel: 1,540
3. IBD gene panel: 489

There was an overlap of 297 genes between the autoimmune panel and the curated list of IBD genes. All genes in these panels are included in the ‘All available genes’ list.

Random forest classifier results

The best IBD subtype classification performance was achieved using the autoimmune gene panel, with an AUROC of 0.68 on the test dataset. The IBD panel achieved an AUROC of 0.61, and all available genes attained an AUROC of 0.58 (Figure 2, Table 2). Hyperparameter tuning results for each classifier can be found in Supplementary Tables 1-3. Regardless of the gene panel used for classification, the *NOD2* gene is the top discriminator of CD and UC. When comparing the results of the classifier using the autoimmune gene panel to the IBD gene panel classifier, both were able to identify UC patients (sensitivity 0.68 for both classifiers). However, the IBD gene panel classifier performs poorly in identification of CD patients in comparison to the autoimmune gene panel (sensitivity 0.46 and 0.63, respectively). Aside from *NOD2*, only the *GC* gene, coding for a Vitamin D binding protein, appears in multiple ML models: the all genes classifier, and the IBD gene panel classifier.

Relative gene importance

The GenePy score distribution has a similar pattern for many genes, with a skew towards a peak at or near zero, reflecting that most individuals have either no variants or very few variants imparting a minimal burden of pathogenic variation, and only a minority demonstrate high scores. This results in skewed distributions with long tails, and often these tails are longer for patients with CD (Figure 3A). The SHAP values generated on the autoimmune gene panel classifier (Figure 3B) also show that for most genes, a low GenePy score contributes towards UC classification, and a high score to CD classification (a positive SHAP value means the gene contributes to the positive class – CD). This is particularly evident for *NOD2,* which has a clear separation in the SHAP value associated with high and low scores. The exceptions to low gene pathogenicity burden (small GenePy score) contributing to a classification of UC, are the genes *IL31RA*, *NRP1*, and *LRP1*. Overall, the SHAP values associated with the top discriminant genes, along with the individual gene importance values for the 719 genes that contribute to classification with the autoimmune gene panel indicates that each gene makes a small contribution to the classification of patients. GenePy score distributions and SHAP values for the all genes and IBD gene panel classifiers shown in Supplementary Figure 1 and 2, respectively (all feature importance values for all ML models available in Supplementary File 3).

*NOD2* as a standalone discriminator between CD and UC

The iterative Fisher’s exact test resulted in an optimised *NOD2* GenePy cut-off of value in the training cohort of 0.2798 for differentiation between CD (above 0.2798) and UC (below 0.2798). AUROC analysis in this data set demonstrated *NOD2* only was able to differentiate between CD and UC with an AUC of 0.61, demonstrating poorer classification ability compared to the top performing ML classifiers. Applying the cut-off value of 0.2798 to a testing set of data demonstrated statistical significance (ꭓ2) to predict CD vs UC using NOD2 alone, p=0.003. Supplementary figure 3.

**DISCUSSION**

Here, we employed a supervised ML algorithm, random forest, to classify IBD patients by subtype, using their whole exome sequencing data summarised into GenePy scores. We demonstrate an AUROC of 0.68 on the test dataset utilising an autoimmune gene panel, which outperformed an IBD gene panel and a classifier using all available genes. This model also out-performs a classifier based on *NOD2* only, although *NOD2* was the most discriminant gene across all classifiers. The current understanding of the genetic drivers of CD indicate that *NOD2* has a significant role in risk of disease, and is perhaps casual for a number of patients.34,35 The autoimmune gene panel classifier outperforming the IBD gene panel suggests that some genes that are currently associated with other autoimmune diseases may also contribute to IBD aetiology. For example, *WDFY4* is present as a top discriminant gene for CD and UC. Previously, this gene had been shown to be associated with systematic lupus erythematosus, and not CD or UC in a GWAS meta-analysis of risk loci associated with autoimmune diseases.36 However, Figure 3A shows clear differences in the tail of the *WDFY4* GenePy score distribution, and high scores in this gene contributing to a classification of CD. This indicates some rare variation is present in a subset of CD, and potentially rare enough to not be detected in GWAS. Further insight into the autoimmune panel genes identified by the random forest classifier could be gained through gene set enrichment analysis. As demonstrated by the SHAP values shown, most genes provide small contributions towards classification of patients as each subtype. This is consistent with the complex, polygenic nature of IBD pathology.

In general, utilising feature selection to reduce the dimensionality of the data, alongside hyperparameter tuning, leads to a more robust and generalisable ML model. A limitation of this pipeline is that these processes were performed sequentially, rather than optimising the parameters and hyperparameters of the ML model together. This process would have been computationally intensive, especially when utilising all genes for classification, which is why the pipeline was constructed in this way. Another limitation of the model is that only patients with European ancestry were included, meaning the results here may not be applicable across all genetic ancestries. This pre-processing step was performed, along with the removal of patients that were related, so that no genomic signals were introduced into the model that were unrelated to IBD subtypes, that could potentially cause model bias

Earlier work that has utilised WES data and ML for IBD was published in response to the Critical Assessment of Genome Interpretation (CAGI) challenge, for classification of CD patients and controls. These datasets were relatively small, and one of the three datasets is known to have batch effects.37 More contemporary work has seen WES data summarised into gene mutational burden scores, again for classification of CD patients and controls. Wang et al. utilised variant consequence (e.g. indel, missense) and zygosity to construct scores38, while Raimondi et al used variant consequence, and weighted genes according to the number of publications associating that gene with IBD. 39 Here, we used WES data for a more clinically applicable disease subtype classifier. Another advantage of our classifier is the disease burden scoring algorithm GenePy, which integrates highly relevant information to a variant’s predicted impact (allele frequency, zygosity, and predicted pathogenicity).

Comparison of the merits of our novel methodology to polygenic risk scores (PRS) is important, with prediction of disease being possible with previous PRS40. From a mathematical perspective we employ a non-linear approach (compared to the linear relationships established by PRS), which allow identification of more complex relationships between data and outcome. Perhaps the biggest advantage of our novel ML approach is the inclusion of whole exome sequencing data, and the ability to include rare variation into disease prediction models. Furthermore, including a per gene deleteriousness metric as the input for the model provides significantly more biological insight than PRS, with specific genes being discriminating features, rather than ‘risk’ loci. Including contemporary sequencing data that encompasses all variants, regardless of minor allele frequency or variant type, is clearly important. Further refinement of our model could occur with whole genomic sequencing whereby promotor/regulatory/splicing control for each gene is included in the in the per gene deleteriousness metric.

In the random forest classifier using the autoimmune gene panel and IBD gene panel, it was interesting to note that while *NOD2* was a top discriminator, the classifier was most sensitive to the UC class, indicating a low *NOD2* score was more associated with a diagnosis of UC, compared to a high score being associated with CD. A potential theory here is that although there are CD patients with high GenePy scores in genes, that the more consistent pattern identified by the random forest classifier is the *lack* of genomic variation in these genes in UC patients. There are potentially many combinations of genomic variation that cause CD, and at this sample size the random forest classifier may be limited in identifying these genomic subgroups and assigning the correct subtype label. This heterogeneity within subtypes has previously been shown with unsupervised learning using endoscopy and histology data.16 *NOD2* has previously been identified as the strongest genomic driver of Crohn’s disease, and has more recently been demonstrated to be useful as a genomic biomarker of stricturing disease29,41. The ability of *NOD2* to distinguish phenotypes appears to be considerable but it remains only a part of the genomic complexity of disease.

The ML classification performance achieved here is promising, considering that genomic variation is one of many factors associated with IBD aetiology. In addition, WES data is sparse and highly dimensional due to the 135,867 exonic variants in the dataset, each of which is only present in a subset of the cohort. Therefore, transformation of the dataset into GenePy scores to reduce both data sparsity and dimensionality is crucial. Larger datasets may be one avenue for the improvement of disease subtype-based classifiers. There are clearly many combinations of genetic variation that can lead to the development of IBD. In this study we include both adult and paediatric patients, data would indicate that the genomic architecture remains consistent regardless of age of onset (with the exception of monogenic forms of IBD) but the effect size of genomic variation is higher in paediatric-onset disease42 [7]. More data could enable better detection of the different combinations of genomic pathogenicity burden by ML algorithms that can lead to each IBD subtype. The overwhelming majority of IBD genetic studies have been conducted on Caucasian populations from North America or Europe, meaning the genes associated with IBD are also population specific. Classification models trained on these specific data are also specific to the population that the model was trained on. A key advantage of ML modelling is that the model algorithm is naïve to which genes have been previously associated with IBD, meaning that understudied populations could easily have models constructed, if the genomic data were available. Datasets such as UK Biobank will be valuable for stratifying patients based on their genomic signal. Of course, there may be a proportion of patients for whom classification based on subtype and WES data is not possible, given the evident genetic heterogeneity. For some patients, their case of disease may be rare monogenic or digenic variation. Other patients may have specific, familial patterns of genomic variation. A highly heterogeneous population partly explains why the ML model AUROC is only modestly good. It may be the case that there is a limit on the AUROC it is possible to achieve for subtype classification using genomic data alone. Therefore, unsupervised clustering may provide better insight into patient subgroups, where disease is driven by shared molecular mechanisms. This approach is more suitable for genomic signal discovery. In the case of driving forward ML classifiers, a narrower focus on specific IBD complications or phenotypes, such as the stricturing or penetrating endotypes in CD patients, may result in better stratification. These specific pathologies may have less variation in their genetic basis. Further, such prognostic models may prove even more useful to clinicians than subtype predictions.

**FUNDING STATEMENT**

This study was supported by the Institute for Life Sciences, University of Southampton, and the National Institute for Health Research (NIHR) Southampton Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

**ACKNOWLEDGEMENTS**

The authors would like to thank Rachel Haggarty for assistance with recruitment and management of the genetics of PIBD study database and Nicola Graham for their assistance with management and extraction of DNA samples. We would like to acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the processing of whole exome sequencing data. Above all, we would like to thank the patients, and their families.

**DATA AVAILABILITY STATEMENT**

The dataset analysed in this study is available through direct collaborative agreement, in line with the informed consent gained from all participants.

**AUTHOR CONTRIBUTION STATEMENT**

ISS and GC processed and transformed the whole exome sequencing data. GC and JJA contributed the in-house IBD gene panel. ISS and EM constructed the machine learning pipeline. ISS performed machine learning analysis and interpretation. JJA and RMB provided clinical interpretation. RMB and SE supervised the research. All authors contributed to the drafting and/or revision of the manuscript.

**CONFLICT OF INTEREST**

All authors declare that they have no conflicts of interest to disclose

**REFERENCES**

1. Levine A, Griffiths A, Markowitz J, et al. Pediatric modification of the montreal classification for inflammatory bowel disease: The paris classification. Inflamm Bowel Dis 2011;17:1314-21.
2. Zaharie R, Tantau A, Zaharie F, et al. Diagnostic delay in romanian patients with inflammatory bowel disease: Risk factors and impact on the disease course and need for surgery. Journal of Crohn's & colitis 2016;10:306-14.
3. Moon CM, Jung SA, Kim SE, et al. Clinical factors and disease course related to diagnostic delay in korean crohn's disease patients: Results from the connect study. PLoS One 2015;10:e0144390.
4. Lamb CA, Kennedy NA, Raine T, et al. British Society of Gastroenterology consensus guidelines on the management of inflammatory bowel disease in adults. Gut. 2019;68:s1–106. doi:10.1136/gutjnl-2019-318484
5. Ricciuto A, Fish JR, Tomalty DE, et al. Diagnostic delay in canadian children with inflammatory bowel disease is more common in crohn's disease and associated with decreased height. Arch Dis Child 2018;103:319-26.
6. Hugot J-P, Chamaillard M, Zouali H, et al. Association of nod2 leucine-rich repeat variants with susceptibility to crohn's disease. Nature 2001;411:599-603.
7. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 2012;491:119-24.
8. Liu JZ, van Sommeren S, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nature genetics 2015;47:979-86.
9. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of gwas loci identifies independent rare variants associated with inflammatory bowel disease. Nature genetics 2011;43:1066-73.
10. Peplow M. The 100 000 genomes project. BMJ 2016;353:i1757.
11. The “all of us” research program. New England Journal of Medicine 2019;381:668-76.
12. Sloutsky R, Jimenez N, Swamidass SJ, Naegle KM. Accounting for noise when clustering biological data. Briefings in Bioinformatics 2012;14:423-36.
13. Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence 1997;97:245-71.
14. Stafford IS, Kellermann M, Mossotto E, et al. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. npj Digital Medicine 2020;3:30.
15. Stafford IS, Gosink MM, Mossotto E, Ennis S, Hauben M. A systematic review of artificial intelligence and machine learning applications to inflammatory bowel disease, with practical guidelines for interpretation. Inflammatory Bowel Diseases 2022.
16. Mossotto E, Ashton JJ, Coelho T, et al. Classification of paediatric inflammatory bowel disease using machine learning. Scientific reports 2017;7:2427.
17. Levine A, Koletzko S, Turner D, et al. Espghan revised porto criteria for the diagnosis of inflammatory bowel disease in children and adolescents. J Pediatr Gastroenterol Nutr 2014;58:795-806.
18. Lamb CA, Kennedy NA, Raine T, et al. British society of gastroenterology consensus guidelines on the management of inflammatory bowel disease in adults. Gut 2019;68:s1.
19. Li H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. ArXiv 2013;1303.
20. Van der Auwera GA, Carneiro MO, Hartl C, et al. From fastq data to high confidence variant calls: The genome analysis toolkit best practices pipeline. Current protocols in bioinformatics 2013;43:11.0.1-.0.33.
21. Rentzsch P, Schubach M, Shendure J, Kircher M. Cadd-splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Medicine 2021;13:31.
22. Carson AR, Smith EN, Matsui H, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC bioinformatics 2014;15:125-.
23. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581:434-43.
24. Mossotto E, Ashton JJ, O’Gorman L, et al. Genepy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. BMC Bioinformatics 2019;20:254.
25. Fuentes Fajardo KV, Adams D, Program NCS, et al. Detecting false-positive signals in exome sequencing. Human Mutation 2012;33:609-13.
26. Pedersen BS, Quinlan AR. Who's who? Detecting and resolving sample anomalies in human DNA sequencing studies with peddy. The American Journal of Human Genetics 2017;100:406-13.
27. Qi Y. Random forest for bioinformatics. In, 2012.
28. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning with applications in r: Springer-Verlag New York; 2013.
29. Ashton JJ, Cheng G, Stafford IS, et al. Prediction of Crohn’s Disease Stricturing Phenotype Using a NOD2-derived Genomic Biomarker. Inflamm Bowel Dis 2022;:1–11. doi:10.1093/IBD/IZAC205
30. Ashton JJ, Boukas K, Davies J, et al. Ileal transcriptomic analysis in paediatric Crohn’s disease reveals IL17- and NOD- signalling expression signatures in treatment-naïve patients and identifies epithelial cells driving differentially expressed genes. J Crohns Colitis Published Online First: 24 November 2020. doi:10.1093/ecco-jcc/jjaa236
31. Bolton C, Smillie CS, Pandey S, et al. An Integrated Taxonomy for Monogenic Inflammatory Bowel Disease. Gastroenterology Published Online First: November 2021. doi:10.1053/J.GASTRO.2021.11.014
32. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in python. the Journal of machine Learning research 2011;12:2825-30.
33. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems Long Beach, California, USA: Curran Associates Inc., 2017: 4768–77.
34. Horowitz JE, Warner N, Staples J, et al. Mutation spectrum of nod2 reveals recessive inheritance as a main driver of early onset crohn’s disease. Scientific Reports 2021;11:5595.
35. Ashton JJ, Mossotto E, Stafford IS, et al. Genetic sequencing of pediatric patients identifies mutations in monogenic inflammatory bowel disease genes that translate to distinct clinical phenotypes. Clin Transl Gastroenterol 2020;11:e00129-e.
36. Ramos PS, Criswell LA, Moser KL, et al. A comprehensive analysis of shared loci between systemic lupus erythematosus (sle) and sixteen autoimmune diseases reveals limited genetic overlap. PLOS Genetics 2011;7:e1002406.
37. Giollo M, Jones DT, Carraro M, et al. Crohn disease risk prediction-best practices and pitfalls with exome data. Hum Mutat 2017;38:1193-200.
38. Wang Y, Miller M, Astrakhan Y, et al. Identifying crohn's disease signal from variome analysis. Genome Med 2019;11:59.
39. Raimondi D, Simm J, Arany A, et al. An interpretable low-complexity machine learning framework for robust exome-based in-silico diagnosis of crohn's disease patients. NAR Genom Bioinform 2020;2:lqaa011.
40. Cleynen I, González JR, Figueroa C, et al. Genetic factors conferring an increased susceptibility to develop Crohn’s disease also influence disease phenotype: Results from the IBDchip European project. Gut 2013;62:1556–65. doi:10.1136/gutjnl-2011-300777
41. Ashton JJ, Seaby EG, Beattie RM, et al. NOD2 in Crohn’s disease- unfinished business. J Crohns Colitis Published Online First: 25 August 2022. doi:10.1093/ECCO-JCC/JJAC124
42. Graham DB, Xavier RJ. Pathway paradigms revealed from the genetics of inflammatory bowel disease. Nature 2020;578:527–39. doi:10.1038/s41586-020-2025-2

**FIGURE LEGENDS**

**Figure 1** Machine learning pipeline. Workflow shows the input data and gene panels, and corresponding pre-processing and data transformation (scaling). Feature selection with the linear support vector classifier (SVC) was performed within a 10-fold cross validation scheme before proceeding to identify the best hyperparameter values with Bayesian optimisation and nested cross-validation (7-fold outer, 5-fold inner). GenePy scores of genes selected by feature selection, and optimal hyperparameter values were input for training the random forest. Machine learning metrics including AUROC, sensitivity and specificity were collected from the random forest’s performance in classification of the testing data.

**Figure 2** AUROC curves for each gene panel used on the testing data.

**Figure 3** Gene GenePy score distributions, and their contributions to the random forest model. A) Distributions of GenePy scores of the top 10 genes in the classifier that utilised the autoimmune gene panel, grouped by IBD subtype. B) SHAP values representing GenePy scores contributions to classification by random forest. In this context the value on the x-axis demonstrates the contribution of that gene to the prediction, with the colour of that point demonstrating the directionality of that contribution related to the positive class (Crohn’s disease). *NOD2* with low values (represented by blue) are highly important for prediction that an individual does not have Crohn’s disease (negative SHAP value). Similarly high *NOD2* values (pink) are important for classification as Crohn’s disease, but this is applicable to fewer cases.

**TABLES**

**Table 1** Number of individuals per IBD subtype in the training and testing data for machine learning. Includes the number and corresponding percentage of individuals in each dataset and disease subtype with paediatric disease onset (here defined as an age at diagnosis of <18 years of age).

|  |  |  |  |
| --- | --- | --- | --- |
|  | CD (no. paediatric onset, %) | UC (no. paediatric onset, %) | Total (no. paediatric onset, %) |
| Training Dataset | 244 (133, 54.5%) | 244 (125, 51.2%) | 488 (258, 52.9%) |
| Testing Dataset | 356 (201, 56.5%) | 62 (32, 51.6%) | 418 (233, 55.7%) |
| Total | 600 (334, 55.7%) | 306 (157, 51.3%) | 906 (491, 54.2%) |

Table 2 Random forest classifier results. Includes the number of genes selected by linear SVC feature selection, top 10 most discriminant genes determined during training, and machine learning assessment metrics. All machine learning metrics are from random forest performance on the test dataset.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ALL GENES** | | | | | **AUTOIMMUNE PANEL GENES** | | | | | **CD vs UC – IBD PANEL GENES** | | | | |
| No. Features | 1,187 | | | | No. Features | 719 | | | | No. Features | 411 | | | |
|  | Precision | Recall | Specificity | F1 |  | Precision | Recall | Specificity | F1 |  | Precision | Recall | Specificity | F1 |
| CD | 0.87 | 0.58 | 0.50 | 0.69 | CD | 0.92 | 0.63 | 0.68 | 0.75 | CD | 0.89 | 0.46 | 0.68 | 0.61 |
| UC | 0.17 | 0.50 | 0.58 | 0.26 | UC | 0.24 | 0.68 | 0.63 | 0.36 | UC | 0.18 | 0.68 | 0.46 | 0.28 |
| Average | 0.77 | 0.57 | 0.51 | 0.63 | Average | 0.82 | 0.64 | 0.67 | 0.69 | Average | 0.79 | 0.49 | 0.65 | 0.56 |
| AUROC | 0.57 | | | | AUROC | 0.68 | | | | AUROC | 0.61 | | | |
| Top 10 Genes | *NOD2, GC, EPB41L4A, ASPM, LAMA1, COL4A3, DNAH17, TUBB3, MYO18B, VWDE* | | | | Top 10 Genes | *NOD2, ATM, JAG1, E2F4, NRP1, IL31RA, LRP1, DNAH12, WDFY4, HHAT* | | | | Top 10 Genes | *NOD2, GC, NFATC1, CELSR3, GALC, DOCK8, ELF1, ITGAL, NPC1, CYBA* | | | |

**FIGURES**





