**A systematic review of artificial intelligence and machine learning applications to inflammatory bowel disease, with practical guidelines for interpretation**

Imogen S. Stafford MSci1,2,3\*, Mark M. Gosink PhD4\*, Enrico Mossotto PhD1, Sarah Ennis PhD1Ŧ, Manfred Hauben MD, MPH4,5Ŧ

1. Human Genetics and Genomic Medicine, University of Southampton, Southampton, UK
2. Institute for Life Sciences, University Of Southampton, Southampton, UK
3. NIHR Southampton Biomedical Research, University Hospital Southampton, Southampton, UK
4. Pfizer Inc, New York, NY, USA
5. NYU Langone Health, Department of Medicine, New York, NY, USA

\* Denotes joint first author

Ŧ Denotes joint last author

**Address for correspondence**

Sarah Ennis

Department of Human Genetics and Genomic Medicine,

University of Southampton,

Southampton,

UK

s.ennis@soton.ac.uk

**Word Count**:3658

**Declaration:** The authors declare that they have no financial disclosures to make, and there are no conflicts of interest to report.

**SUMMARY**

This systematic review identifies the application of machine learning to inflammatory bowel disease, concerning popular methods, tasks and data types in the field. Recent trends are identified, and aspects of these method’s pipelines, including sample size and cross-validation, are discussed.

**ABSTRACT**

Background: Inflammatory bowel disease (IBD) is a gastrointestinal chronic disease with an unpredictable disease course. Computational methods such as machine learning (ML) have the potential to stratify IBD patients for the provision of individualised care. The use of ML methods for IBD was surveyed, with an additional focus on how the field has changed over time.

Methods: A systematic review was conducted through a search of MEDLINE and Embase databases, with the search structure (“machine learning” OR “artificial intelligence”) AND (“Crohn\* Disease” OR “Ulcerative Colitis” OR “Inflammatory Bowel Disease”), searched 6th May 2021. Exclusion criteria: studies not written in English, no human patient data, publication before 2001, studies that were not peer reviewed, non-autoimmune disease comorbidity research and record types that were not primary research.

Results: 78 (of 409) records met the inclusion criteria. Random forest methods were most prevalent, and there was an increase in neural networks, mainly applied to imaging datasets. The main applications of ML to clinical tasks were diagnosis (18/78), disease course (22/78) and disease severity (16/78). The median sample size was 263. Clinical and microbiome-related datasets were most popular. 5% of studies used an external dataset after training and testing for additional model validation.

Discussion: Availability of longitudinal and deep phenotyping data could lead to better modelling. ML pipelines considering imbalanced data, and feature selection only on training data will generate more generalisable models. ML models are increasingly being applied to more complex clinical tasks for specific phenotypes, indicating progress towards personalised medicine for IBD.

**Key Words**: artificial intelligence, machine learning, inflammatory bowel disease

**INTRODUCTION**

Inflammatory bowel disease (IBD) is an umbrella term for a set of chronic diseases, of which there are two main subtypes: Crohn’s disease (CD) and ulcerative colitis (UC). The global prevalence of IBD increased to 84.3 per 100,000 by 2017, and with it comes a greater burden to patients and health services [1]. Due to a number of factors contributing to its aetiology, IBD disease course is highly variable. Patients can experience a mild disease, or a severe, refractory disease requiring many interventions. A patient’s disease course is often unclear at diagnosis.

There has been a relative explosion in the use of artificial intelligence and machine learning (ML) techniques for complex diseases, after the success of these algorithms in fields like oncology [2]. Unlike traditional statistical techniques, ML infers patterns from data, allowing model application to unseen cases. Key concepts for this field are included in Box 1, and a further breakdown of ML terms, metrics and methods are detailed elsewhere [3, 4]. For IBD, ML has the potential to improve patient care at every stage of their disease course through prediction modelling: from a quick subtype diagnosis so appropriate treatment can be identified, to assessing disease activity and identifying those patients more likely to develop complications and require surgery. For clinicians, this potential is exciting, but comes with many questions about which ML methods may be successful. Here, practical guidelines are provided to guide interpretation of current and future research in this field (Appendix 1). Although this systematic review centres around applications to IBD, these are general guidelines for ML interpretation.

|  |
| --- |
| Box 1 – Key ConceptsArtificial Intelligence: methods that enable computers to mimic human intelligence.Machine Learning: methods that infer patterns from data to perform a specific task, usually classification or regression.Deep Learning: neural network based approaches that enable machines to train themselves to perform tasks.Supervised Learning: the ML model learns patterns in data, and associates this information with an already present label. The model applies can then apply this learning to new data, and predict these labels.Unsupervised Learning: the ML model identifies patterns and clusters the data in a way that explains the data structure (not according to labels).Feature Selection: a collection of methods that reduce the dimensionality of a dataset, such that ML is performed on a subset of the most informative variables for the task.Cross Validation: a method which can reduce the overfitting of ML models, meaning the results will generalise well to new data. During training the data is split into k folds, and the ML model trained on k-1 folds. The model performance is tested on the final fold, and the process repeated so each fold becomes the test fold exactly once.  |

In a previous, broader systematic review of artificial intelligence and ML applications to autoimmune diseases [3], a number of popular methods and applications were identified, and the research assessed guided some recommendations for the field. In addition, other systematic reviews have been published commenting on this area, including Tontini et al.’s review of artificial intelligence for gastrointestinal endoscopy [5], and Nguyen et al.’s study on machine learning for diagnosis and prognosis in IBD [6]. The aim of this systematic review was to assess common data types, applications and methods in the field of ML for IBD, and evaluate changes in the field over the past few years. The broad scope of this review allows for the assessment of trends, and the recording of the full range of ML applications to IBD.

**METHODS**

**Literature Search**

An electronic literature search was performed using two databases available through OvidSP: MEDLINE(R) and Epub Ahead of Print, In-Process, In-Data-Review & Other Non-Indexed Citations and Daily 1946, and Embase 1974. The literature search was completed in May 2021. Search terms were combined using Boolean operators as follows: (“machine learning” OR “artificial intelligence”) AND (“Crohn\* Disease” OR “Ulcerative Colitis” OR “Inflammatory Bowel Disease”). Any record where these search terms were identified in the title, abstract and/or subject headings would appear in the list of records (last search 6th of May 2021).

**Inclusion and Exclusion Criteria**

This systematic review sought to expand and better characterise a subsection of a previous review of ML in autoimmune disease, therefore the same criteria was employed. Studies that applied ML to IBD, or a subtype of IBD, were included. Studies that used ML for analysis of complications that arise from IBD were also included. Studies that were not written in English, were published prior to 2001, that did not use real human patient data, were not peer reviewed, or were not original research papers were also excluded. Therefore, the following publication types (as labelled by OvidSP) were not assessed during screening: conference abstracts, conference review, editorial, erratum, journal article comment, journal article review, letter, letter comment, note and review. The abstract of each study was assessed by two reviewers independently for inclusion in the systematic review. The full text was assessed where a decision on inclusion could not be made based on the abstract, and a consensus reached by the two reviewers. The following data items were collected for each study that met the criteria: the task ML was applied to, the type of ML (supervised or unsupervised), all ML algorithms trialled by the researchers, the best performing ML algorithm, sample size, clinical population (IBD, UC or CD), data type, the best results achieved, whether a training and testing split was used, if other cross-validation was used, if the model was applied to independent test data, and the year of publication. This systematic review conforms to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards [7].

**Graphical Representations of Data**

Articles were graphically summarized in Sunburst or Pie-chart diagrams using a custom R script utilizing the plotly library [8, 9]. The R scripts can be downloaded from Github (github.com/isstafford/review\_ml\_ibd\_2021). Briefly, all articles were classified according to the machine-learning approach used (Method), the type of information being analysed (Data-Type) and the outcome which is being predicted (Task). These categories and the sorting of studies into them were agreed by all authors. The R script counts unique titles under each level of the three-class hierarchy and the results displayed in a sunburst diagram. The inner most ring represents the highest level of the hierarchy whilst the outer most ring being the individual articles. Since some articles discuss multiple methods, tasks or outcomes, they may be represented more than once on the diagram.

Graphical summaries of sample sizes and uses of ML method types over time were generated using ggplot2 in R [10]. For the sample size graphical summary, the ML method was counted as used if it was recorded as a method in the research paper, even if the ML method did not generate the optimal model. ML methods were sorted into type groups (e.g. ridge regression and logistic regression were both included under “regression”). Multiple methods from the same type group, within the same paper were counted once to avoid skewing the data. In cases where papers investigated multiple classification problems with different sample sizes, each classification problem counted as a separate entry. All ML method groups with sufficient data for a boxplot (n ≥ 5) were included in the visualisation. The same ML method groups were plotted for the use of ML types over time.

**The Role of the Funding Source**

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all data in the study and had final responsibility for the decision to submit for publication.

**RESULTS**

Initially, 409 records were identified and 135 records were subsequently removed as duplicates. Applying the study criteria regarding original research articles, year of publication and language removed 153 entries. Of the remaining 121 screened articles, 33 were excluded after assessing the abstract against the inclusion and exclusion criteria, and a further nine were excluded after a full-text read (Figure 1). A technical analysis of the ML applications in these studies is outside the scope of this review. Here, summary statistics are provided regarding popular methods, applications and data; the sample sizes used and when cross-validation was implemented; and trends in ML usage in recent years. The chosen ML models, and data types used for each type of task is detailed in Table 1.

Of 78 studies included in the systematic review, the majority used supervised ML, with 4 papers employing unsupervised methods [11-14], and 5 utilising both supervised and unsupervised ML [15-19] for varied clinical applications. Many papers trialled different ML methods before selecting the optimal one, and some researchers implemented ML for multiple IBD applications. Three main clinical application areas were identified: diagnosis (23%) [15, 20-36], disease course (28%) [15, 30, 37-56] and disease severity (21%) [19, 57-71]. Diagnosis classification tasks involved differentiating IBD patients (or one subtype) from controls. Studies of disease course examined relapse, remission, and surgery ML classifiers. Disease severity studies sought to predict patients’ IBD activity, or those who may develop complications. The most prevalent method implemented was random forest (47%), with regressions, neural networks and support vector machines also used regularly (31%, 28% and 27%, respectively. Percentages here sum to over 100% as multiple methods were trialled by one study in many instances). Other tree-based methods were used by 22% of studies used tree-based methods (13% boosting with trees, 9% decision trees). Clinical data (41%), and data related to the microbiome (23%) were the most commonly used in ML modelling. The median sample size, not including external validation data sets that were additional to usual training and testing data, was 263 (range 12–7,400,000). A breakdown of sample sizes per ML method used can be viewed in Figure 2. Validation data sets in addition to the expected training and testing sets were used in 5% of studies [12, 32, 41, 72]. Another seven studies trained their models with cross-validation on one dataset, and tested their method on an external, independent dataset [23, 29, 45, 59, 61, 66, 73]. CD data (only) was used in 27 studies [12, 17, 19, 26-29, 32, 37-39, 41, 44, 47-50, 58, 60, 63, 65, 74-79], and UC data (only) was used in 15 studies [25, 40, 42, 46, 52, 55, 59, 61, 62, 64, 68-70, 80, 81], with the remainder (n=36) using a mix of CD and UC data, or IBD data as one class [11, 13-16, 18, 20-24, 30, 31, 33-36, 43, 45, 51, 53, 54, 56, 57, 66, 67, 71-73, 82-88]. Half of the research using UC-only data focused on predicting disease activity with endoscopy data, while the aims of ML classifications on CD data were varied. A breakdown of the method and classification task can be found in Figure 3, which can be customised here (isstafford.github.io/review\_ml\_ibd\_2021/), and more details regarding each study included in this review is found in Supplementary Table 1.

The literature search assessed in the previous systematic review was completed on the 18th of December 2018, therefore comparisons were made between studies published before and during 2018, and those published from 2019 to the literature search date. 53 papers have been published from 2019 to May 2021. Here, if a publication is published online, and subsequently printed in a different year, the first publication date is used. Since the end of 2018, there has been a rapid expansion in the use of neural networks (a deep learning method) for IBD, with 21 studies trialling this method on their data from 2019 onwards, in comparison to one study prior to this. This increase coincides with more imaging datasets (4% 2007-2018, 18% 2019-2021), specifically colonoscopy data, and the majority of neural networks were applied to this data type. Support vector machine, random forest and regression based methods were popular during both time periods (year on year breakdown of ML method group use, Figure 4). More studies utilised two data types in 2019-2021 (8% versus 17%), almost always combining clinical data with another data type. The median sample size of studies has not increased in recent years (N=273 2007-2018, N=257.5 2019-2021). Diagnosis has continued to be a popular ML application, but prior to 2019, investigating treatment response was more popular (24% versus 1.8%), and exploring classification tasks connected to disease course is now the most popular application (12% versus 35.8%).

**DISCUSSION**

The increased use of ML methods for IBD documented here is a demonstration of the wider interest in artificial intelligence for healthcare. Due to the heterogeneity of ML model workflows, data types and reported metrics, it was not possible to ascertain any superior approaches. It is possible that some studies may have been excluded from the review, as Medical Subject Headings (MeSH) were not utilised in the search strategy. However, when the ML subject heading was exploded, the only algorithm specified as a search term was “support vector machine”, which could have biased the search strategy towards only identifying additional papers that used this classifier. An additional limitation was the search of only two databases, as the systematic search focused on capturing models with clinical application. An assessment of the risk of bias was not performed, as there is no clear equivalent of PROBAST (Prediction model Risk Of Bias ASsessment Tool) to assess ML modelling. The construction of a tool that could assess potential ML pipeline bias would be beneficial for the transition of models into clinical settings. Minimising bias in modelling, and creating generalisable models go hand in hand. Ways to assess the generalisability of ML models are addressed in the remainder of the discussion, and in the appendix.

There is a clear dominance of tree-based methods: one or a combination of random forests, decision trees and tree-based boosting methods were implemented by 55% of studies. This is potentially due to decision trees being highly interpretable, with tree boosting and random forest preventing overfitting of this model type. Random forests are also well known as an ML algorithm that can leverage non-linear relationships. This popularity is not inherently a drawback, however a lack of comparison of different ML methods, or a lack of reporting of this comparison in studies, may make developing ML models for clinical application more challenging.

Overall, there was good reporting of a range of informative metrics in these studies, which was particularly important given many of their data sets had imbalanced classes. In cases where imbalanced data is used, a high accuracy score can mask poor prediction of the minority class. While some studies sought to correct these imbalances with algorithm weighting [82], or oversampling of the minority class [48, 64], some researchers did not explicitly address this. Although imbalanced datasets may be representative of the patient population, it is important to consider whether enough samples from both classes were present in ML algorithm training, such that accurate predictions can be made for both the majority and minority class. Another potential ML pipeline issue discovered here was a use of feature selection on the whole dataset, rather than just the training set. With this workflow, there is a danger that information regarding features in the test set leaks into the training set. Improvements to ML pipelines can only benefit patients, as more robust workflows allow the identification of the most successful models.

While an increase in the use of external datasets in the expected workflow of training and testing on one dataset, and validating on an independent dataset was not observed, other interesting approaches were employed, showing researchers bringing datasets together to extract additional information. Some studies used all of their initial dataset to train their model with cross-validation, and subsequently tested the model on a different external dataset [23, 29, 45, 59, 61, 66, 73]. Others used a type of cross validation called LODO, or leave-one-dataset-out, allowing researchers to utilise many, smaller datasets [27, 84].

The range of overall dataset sizes used in studies was large. Some of the smaller sample sizes used may not have been appropriate for the chosen ML method, although evaluating whether there was sufficient data to construct a classifier can be challenging and the required sample size is contingent on the ML task. There is no standard power calculation available for studies using ML. The sample size required depends on the method used, with algorithms such as neural networks requiring more data. This trend was observed in the systematic review: larger datasets were used in conjunction with neural networks. The number of features used for modelling will also affect the required sample size. More features will generally produce a more complex model, so a larger dataset is necessary. If the ML model has generalised well from training to testing data (or other independent data), this is a good indicator the dataset was sufficient. It is also important to consider how representative the dataset is of the patient population. An ML model may perform well in initial training and testing, but if the dataset is biased in the demographics or phenotypes represented, then the modelling may translate poorly when implemented in clinical settings.

Although diagnosis (classifying controls and IBD patients) is still a popular application in 2021, it was encouraging to see the highest percentage of papers addressing issues surrounding disease course in recent years. This suggests that more longitudinal and deeper phenotyping data is being collected, allowing a move towards more precise and complex classification tasks. The median size of datasets has not grown in recent years. While dataset size is not an indicator of data quality, it is surprising that although we are in the era of “big data”, data set sizes are not increasing at a rate they might be expected to. A potential roadblock in garnering larger data sets for more specific classifiers may be linking up these other data types with phenotyping data. A community effort may be necessary to accumulate sufficient datasets, for more accurate and generalisable ML models, and external validation. Despite the uncontested power of ‘omics datasets in providing a -usually- unbiased representation of a patient profile, detailed clinical information remains fundamental for precise phenotyping and patient stratification. Projects such as UK Biobank [89] have progressed this need for data, but phenotyping can be limited. With this data in place, robust pipelines and models that generalise well, the community takes the next step towards personalised medicine for IBD patients.

**APPENDIX: APPLICATION OF AI TO IBD IN LITERATURE: POINTS-TO-CONSIDER FOR CLINICIANS**

Rare indeed is the clinician who possesses the knowledge, experience and time to master all the intersecting disciplines in health care ML research. Compounding this are the susceptibility of AI in health care to hype cycle effects [90], and a documented deficit in the quality and completeness of reporting in AI in health care papers [91].

The interested but busy clinician has aids at their disposal in the form of guidelines and checklists on reporting requirements and quality. Some are tailored for clinicians, others more broadly to peer-reviewers, users and authors [91-101]. We recommend that whenever possible clinicians obtain these, especially those customized for clinicians [91-93, 100] for self-paced study and to have on hand as an assessment support tool when reading papers.

However methodically going through a checklist, some of which are quite lengthy, item-by-item every time a paper is read, while ideal, may still pose a time challenge for the busy clinician, and realistically it is to be expected that they will often read and assess papers against background knowledge/ situational awareness. Herein we provide a brief exposition on our subjective choice of a subset of critical items for clinicians’ situational awareness in the absence of a checklist. They should also enhance their understanding and use of checklists

**I. Comparing ML Algorithms**

Direct model performance measures include accuracy, recall (sensitivity), precision (positive predictive value), specificity, negative predictive value, area under the curve (AUC) and F1-score. Comparing ML algorithms on these measures across studies is not usually meaningful because of unmatched factors, including specific model implementation, database size, type, structure, and quality, reported performance metrics, and specific application. However, studies have compared multiple ML algorithms and curated benchmarking repositories. These repositories document the application of numerous ML algorithms to multiple diverse data sets, and provide general insights on performance gradients where consistent performance gradients are observed [102-104]. The caveat is these insights may not completely generalise, especially with more modern algorithms, such as neural networks. These methods entail many analytical choices which impact ML architecture, performance, and limitations, and the algorithms are highly situational and operator-dependent. In fact, it has been argued that a fair machine versus machine comparison would have to eliminate operator interaction, making the question unanswerable [103]. Generally, newer ML algorithms can leverage current computational capacity to predict on complex, large, high-dimensional better than older traditional ML algorithms, but not always [102-104].

Importantly, choice of a ML model is not only related to direct performance measures. Indirect performance-related ML characteristics are equally important, but perhaps more difficult to compare across different workflows. These include: algorithm susceptibility to over-fitting, which can lead to difficulties achieving the same results on different data; the transparency of ML model decisions; the time and computational cost of ML algorithms; whether the dataset is static (offline learning) or updating over time (online learning); raw data pre-processing (feature engineering); whether the model is robust to outliers, and if the ML algorithm contains statistical assumptions. Even environmental impacts may come to play a role in algorithm selection. Increasing computational complexity correlates with increased energy consumption and greenhouse emissions. One group of investigators estimated that the CO2 emissions of a full training cycle for a neural network emits approximates the lifetime CO2 emissions of four cars [105].

ML model selection will therefore always entail some trade-offs. A complex algorithm may deliver excellent performance for the task, but at high computational cost, and a limited understanding of how this performance was achieved. All ML algorithms have different strengths and weaknesses. Support vector machines, which find a boundary between different classes of points by optimally separating the closest points from the two classes, are less susceptible to overfitting and data outliers, but very large and noisy datasets present a challenge for the algorithm to extract a meaningful boundary between two classes. In contrast, a properly initialised deep neural network can approximate any complex decision boundary, identifying and exploiting, if not revealing, complex interactions. The constructed model may be highly accurate, but has a propensity to overfit, and sensitivity to initialisation as the cost. For all ML methods, it is important to be aware of any assumptions contained within the model. This is especially important for more assumption-laden traditional ML. However, if those assumptions are met, the models can return a quick, competitive performance. It is therefore apt to consider whether a statistically significant ML performance translates into clinical impact for a specific application, given the totality of direct and indirect measures.

**II. Dataset Quality, Construction, and Labelling**

1. Dataset Quality, Construction and Labelling

Good training, test and validation datasets are foundational. They should have accurate class labels (i.e. true positive and negative instances) that are well defined, and capture the full range of clinical, demographic and practice variables to be generalisable. Clinical expertise is as important as ML expertise to detect unrepresentative data. Training, test and validation sets should originate from independent sources, or the data should be randomly (not manually) split. Care should be taken to prevent data leakage (test data information leaking into training data), causing biased results, poor generalisation, and camouflaged over-fitting. This can occur in data with multiple samples, especially time series, from individual patients, or with data pre-processing or transformation prior to splitting. For example, normalizing a continuous variable in the prior to splitting data by using its global mean and standard deviation, is leaking some information from the test set into the training set.

Clinician rating schemes for diagnostic labelling (e.g. labelling patients in the data set according to IBD disease activity, endoscopic image-based diagnosis) can affect performance, for example assessment by majority vote versus full adjudication on the same reference set can return significantly different error rates [100]. Blinded adjudication by an expert panel provided with sufficient information and time is ideal for subjective labels.

Opportunities for poor representativeness abound. An IBD-specific example is generalising from general/adult-focused IBD data to early-onset IBD, with its unique phenotypes. Dichotomous classification, for example IBD versus healthy controls, is common and does not accurately reflect differential diagnosis in the clinic, pathology lab, radiology reading room, or endoscopy suite, involving multiple diagnostic possibilities, thus generalising poorly to real-world settings. In endoscopy studies, deficient representation can be caused by many factors such as endoscope brands, endoscopic modality (high versus standard resolution white light, chromoendoscopy), operator skill, number of study sites , inclusion/exclusion criteria (e.g. only best archived images, patients with adequate bowel preps). In digital pathology, randomly cropped versus whole slide images presence/absence of standardisation of whole-slide imaging and staining, could return different results [106, 107]. Distributional shifts occur when characteristics and context of contemporary data, such as evolving clinical phenotypes, diverge from the training data, resulting in an outdated representation. Ongoing acquisition and use of training data representing current disease or practice are required [108, 109].

2. Imbalanced Datasets

Imbalanced data is unavoidable, especially as ML algorithms move towards more complex prediction tasks. It may bias performance towards predicting the majority class at the expense of the minority class, and makes overall accuracy an unreliable performance measure, as high accuracy can be achieved by just naively guessing the majority class. The severity of this effect is amplified with smaller, poorly separated, and/or sub-optimally labelled data sets.. Readers should determine if significant imbalance reflects a natural distribution versus an artifact of study execution. It is tempting to just rebalance the data by oversampling the minority class, either with real or synthetic data, or under-sampling the minority class, but if the imbalance is natural, performance on the unrealistically rebalanced data set may diverge from real-world results, which tend to be biased to the minority class. Oversampling should occur after training/testing set splitting to avoid data leakage. If practical, enlarging the data set without distorting the natural proportions is desirable. Other approaches include prioritising the importance, or more heavily penalising false positives or false negatives, according to the specific problem, or finding a feature set in which the classes are more separable. Finally, trying another ML method such as tree-based approach, may be indicated.

Papers with imbalanced data should provide a full suite of direct performance metrics, given the untrustworthiness of overall accuracy in this scenario. Examining the confusion matrix is key to assess class-specific performance, and it serves as the basis for calculating other measures such as precision, recall balanced accuracy and F1 score. For performance measures that are a function of prevalence, such as predictive value, extrapolation from rebalanced data sets to the natural prevalence ratio should be provided.

**ACKNOWLEDGEMENTS**

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.

**REFERENCES**

[1] Alatab S, et al. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990&#x2013;2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet Gastroenterology & Hepatology. 5, 17-30 (2020).

[2] McKinney SM, et al. International evaluation of an AI system for breast cancer screening. Nature. 577, 89-94 (2020).

[3] Stafford IS, et al. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. npj Digital Medicine. 3, 30 (2020).

[4] Fatima M, Pasha M. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications. 09, 1-16 (2017).

[5] Tontini GE, et al. Artificial intelligence in gastrointestinal endoscopy for inflammatory bowel disease: a systematic review and new horizons. Therapeutic advances in gastroenterology. 14, 17562848211017730-17562848211017730 (2021).

[6] Nguyen NH, et al. Machine Learning-based Prediction Models for Diagnosis and Prognosis in Inflammatory Bowel Diseases: A Systematic Review. Journal of Crohn's and Colitis. 16, 398-413 (2021).

[7] Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Medicine. 6, e1000097 (2009).

[8] Team RC. R: A language and environment for statistical computing. (2019).

[9] Inc. PT. Collaborative data science. Montréal, QC: Plotly Technologies Inc.; 2015.

[10] Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York. (2016).

[11] Liu T, et al. Distinct clinical phenotypes for Crohn's disease derived from patient surveys. BMC Gastroenterol. 21, 160 (2021).

[12] Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome. 8, 90 (2020).

[13] Coelho T, et al. Immunological Profiling of Paediatric Inflammatory Bowel Disease Using Unsupervised Machine Learning. J Pediatr Gastroenterol Nutr. 70, 833-840 (2020).

[14] Lerrigo R, et al. The Emotional Toll of Inflammatory Bowel Disease: Using Machine Learning to Analyze Online Community Forum Discourse. Crohn's & Colitis 360. 1, (2019).

[15] Clooney AG, et al. Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. Gut. 70, 499-510 (2021).

[16] Dhaliwal J, et al. Accurate Classification of Pediatric Colonic Inflammatory Bowel Disease Subtype Using a Random Forest Machine Learning Classifier. J Pediatr Gastroenterol Nutr. 72, 262-269 (2021).

[17] Le V, Quinn TP, Tran T, Venkatesh S. Deep in the Bowel: Highly Interpretable Neural Encoder-Decoder Networks Predict Gut Metabolites from Gut Microbiome. BMC Genomics. 21, 256 (2020).

[18] Mossotto E, et al. Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. Scientific reports. 7, 2427 (2017).

[19] Niehaus KE, Uhlig HH, Clifton DA. Phenotypic characterisation of Crohn's disease severity. Annu Int Conf IEEE Eng Med Biol Soc. 2015, 7023-7026 (2015).

[20] Biernacka KB, et al. The value of magnetic resonance enterography in diagnostic difficulties associated with Crohn's disease. Pol J Radiol. 86, e143-e150 (2021).

[21] Volkova A, Ruggles KV. Predictive Metagenomic Analysis of Autoimmune Disease Identifies Robust Autoimmunity and Disease Specific Microbial Signatures. Front Microbiol. 12, 621310 (2021).

[22] Nuzzo A, et al. Expanding the drug discovery space with predicted metabolite-target interactions. Commun Biol. 4, 288 (2021).

[23] Xu C, et al. LightCUD: a program for diagnosing IBD based on human gut microbiome data. BioData Min. 14, 2 (2021).

[24] Manandhar I, et al. Gut microbiome-based supervised machine learning for clinical diagnosis of inflammatory bowel diseases. Am J Physiol Gastrointest Liver Physiol. (2021).

[25] Khorasani HM, Usefi H, Peña-Castillo L. Detecting ulcerative colitis from colon samples using efficient feature selection and machine learning. Sci Rep. 10, 13744 (2020).

[26] Raimondi D, et al. An interpretable low-complexity machine learning framework for robust exome-based in-silico diagnosis of Crohn's disease patients. NAR Genom Bioinform. 2, lqaa011 (2020).

[27] Jiang P, Lai S, Wu S, Zhao XM, Chen WH. Host DNA contents in fecal metagenomics as a biomarker for intestinal diseases and effective treatment. BMC Genomics. 21, 348 (2020).

[28] Romagnoni A, Jégou S, Van Steen K, Wainrib G, Hugot JP. Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. Sci Rep. 9, 10351 (2019).

[29] Wang Y, et al. Identifying Crohn's disease signal from variome analysis. Genome Med. 11, 59 (2019).

[30] Sarrabayrouse G, et al. Fungal and Bacterial Loads: Noninvasive Inflammatory Bowel Disease Biomarkers for the Clinical Setting. mSystems. 6, (2021).

[31] Iablokov SN, et al. Metabolic Phenotypes as Potential Biomarkers for Linking Gut Microbiome With Inflammatory Bowel Diseases. Front Mol Biosci. 7, 603740 (2020).

[32] Douglas GM, et al. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. Microbiome. 6, 13 (2018).

[33] Eck A, et al. Interpretation of microbiota-based diagnostics by explaining individual classifier decisions. BMC Bioinformatics. 18, 441 (2017).

[34] Hübenthal M, et al. Sparse Modeling Reveals miRNA Signatures for Diagnostics of Inflammatory Bowel Disease. PLoS One. 10, e0140155 (2015).

[35] Cui H, Zhang X. Alignment-free supervised classification of metagenomes by recursive SVM. BMC Genomics. 14, 641 (2013).

[36] Forbes JD, et al. A comparative study of the gut microbiota in immune-mediated inflammatory diseases-does a common dysbiosis exist? Microbiome. 6, 221 (2018).

[37] Waljee AK, et al. Assessing Clinical Disease Recurrence Using Laboratory Data in Surgically Resected Patients From the TOPPIC Trial. Crohn's & Colitis 360. 2, (2020).

[38] Stidham RW, et al. The Use of Readily Available Longitudinal Data to Predict the Likelihood of Surgery in Crohn Disease. Inflamm Bowel Dis. (2021).

[39] Udristoiu AL, et al. Deep Learning Algorithm for the Confirmation of Mucosal Healing in Crohn's Disease, Based on Confocal Laser Endomicroscopy Images. J Gastrointestin Liver Dis. 30, 59-65 (2021).

[40] Sakurai T, et al. Mucosal microbiota and gene expression are associated with long-term remission after discontinuation of adalimumab in ulcerative colitis. Sci Rep. 10, 19186 (2020).

[41] Kang EA, et al. Development of a Clinical and Genetic Prediction Model for Early Intestinal Resection in Patients with Crohn's Disease: Results from the IMPACT Study. Journal of clinical medicine. 10, (2021).

[42] Sofo L, et al. New perspectives in the prediction of postoperative complications for high-risk ulcerative colitis patients: machine learning preliminary approach. Eur Rev Med Pharmacol Sci. 24, 12781-12787 (2020).

[43] Shivaji UN, et al. Clinical outcomes, predictors of prognosis and health economics consequences in IBD patients after discontinuation of the first biological therapy. Therap Adv Gastroenterol. 13, 1756284820981216 (2020).

[44] Taylor H, et al. Multiomic features associated with mucosal healing and inflammation in paediatric Crohn's disease. Aliment Pharmacol Ther. 52, 1491-1502 (2020).

[45] Choi YI, et al. Development of Machine Learning Model to Predict the 5-Year Risk of Starting Biologic Agents in Patients with Inflammatory Bowel Disease (IBD): K-CDM Network Study. Journal of clinical medicine. 9, (2020).

[46] Ghoshal UC, Rai S, Kulkarni A, Gupta A. Prediction of outcome of treatment of acute severe ulcerative colitis using principal component analysis and artificial intelligence. JGH Open. 4, 889-897 (2020).

[47] Jones CMA, et al. Bacterial Taxa and Functions Are Predictive of Sustained Remission Following Exclusive Enteral Nutrition in Pediatric Crohn's Disease. Inflamm Bowel Dis. 26, 1026-1037 (2020).

[48] Dong Y, et al. A novel surgical predictive model for Chinese Crohn's disease patients. Medicine. 98, e17510 (2019).

[49] Braun T, et al. Individualized Dynamics in the Gut Microbiota Precede Crohn's Disease Flares. The American journal of gastroenterology. 114, 1142-1151 (2019).

[50] Bottigliengo D, et al. The Role of Genetic Factors in Characterizing Extra-Intestinal Manifestations in Crohn's Disease Patients: Are Bayesian Machine Learning Methods Improving Outcome Predictions? Journal of clinical medicine. 8, (2019).

[51] Waljee AK, et al. Development and Validation of Machine Learning Models in Prediction of Remission in Patients With Moderate to Severe Crohn Disease. JAMA network open. 2, e193721 (2019).

[52] Takenaka K, et al. Development and Validation of a Deep Neural Network for Accurate Evaluation of Endoscopic Images From Patients With Ulcerative Colitis. Gastroenterology. 158, 2150-2157 (2020).

[53] Morell Miranda P, Bertolini F, Kadarmideen H. Investigation of gut microbiome association with inflammatory bowel disease and depression: a machine learning approach [version 2; peer review: 2 approved with reservations]. F1000Research. 7, (2019).

[54] Waljee AK, et al. Predicting Hospitalization and Outpatient Corticosteroid Use in Inflammatory Bowel Disease Patients Using Machine Learning. Inflamm Bowel Dis. 24, 45-53 (2017).

[55] Jain S, et al. Predictors of long-term outcomes in patients with acute severe colitis: A northern Indian cohort study. J Gastroenterol Hepatol. 33, 615-622 (2018).

[56] Firouzi F, et al. A decision tree-based approach for determining low bone mineral density in inflammatory bowel disease using WEKA software. European journal of gastroenterology & hepatology. 19, 1075-1081 (2007).

[57] Dorofeyev AE, Holub SV, Babayeva GH, Аnaniin О. APPLICATION OF INTELLECTUAL MONITORING INFORMATION TECHNOLOGY IN DETERMINING THE SEVERITY OF THE CONDITION OF PATIENTS WITH INFLAMMATORY BOWEL DISEASES. Wiad Lek. 74, 481-486 (2021).

[58] Ungaro RC, et al. Machine learning identifies novel blood protein predictors of penetrating and stricturing complications in newly diagnosed paediatric Crohn's disease. Aliment Pharmacol Ther. 53, 281-290 (2021).

[59] Gutierrez Becker B, et al. Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data. Ther Adv Gastrointest Endosc. 14, 2631774521990623 (2021).

[60] Li X, et al. Development and Validation of a Novel Computed-Tomography Enterography Radiomic Approach for Characterization of Intestinal Fibrosis in Crohn's Disease. Gastroenterology. 160, 2303-2316.e2311 (2021).

[61] Yao H, et al. Fully automated endoscopic disease activity assessment in ulcerative colitis. Gastrointest Endosc. 93, 728-736.e721 (2021).

[62] Gottlieb K, et al. Central Reading of Ulcerative Colitis Clinical Trial Videos Using Neural Networks. Gastroenterology. 160, 710-719.e712 (2021).

[63] Wang J, et al. High circulating elafin levels are associated with Crohn's disease-associated intestinal strictures. PLoS One. 15, e0231796 (2020).

[64] Popa IV, Burlacu A, Mihai C, Prelipcean CC. A Machine Learning Model Accurately Predicts Ulcerative Colitis Activity at One Year in Patients Treated with Anti-Tumour Necrosis Factor α Agents. Medicina (Kaunas). 56, (2020).

[65] Reddy BK, Delen D, Agrawal RK. Predicting and explaining inflammation in Crohn's disease patients using predictive analytics methods and electronic medical record data. Health Inform J. 1460458217751015 (2018).

[66] Biasci D, et al. A blood-based prognostic biomarker in IBD. Gut. 68, 1386-1395 (2019).

[67] Mohapatra S, et al. Wavelet Transform and Deep Convolutional Neural Network-Based Smart Healthcare System for Gastrointestinal Disease Detection. Interdiscip Sci. 13, 212-228 (2021).

[68] Takenaka K, et al. Deep Neural Network Accurately Predicts Prognosis of Ulcerative Colitis Using Endoscopic Images. Gastroenterology. 160, 2175-2177.e2173 (2021).

[69] Bossuyt P, et al. Automatic, computer-aided determination of endoscopic and histological inflammation in patients with mild to moderate ulcerative colitis based on red density. Gut. 69, 1778-1786 (2020).

[70] Maeda Y, et al. Fully automated diagnostic system with artificial intelligence using endocytoscopy to identify the presence of histologic inflammation associated with ulcerative colitis (with video). Gastrointestinal Endoscopy. 89(2), 408-415 (2018).

[71] Menti E, et al. Bayesian Machine Learning Techniques for revealing complex interactions among genetic and clinical factors in association with extra-intestinal Manifestations in IBD patients. AMIA Annu Symp Proc. 2016, 884-893 (2016).

[72] Waljee AK, et al. Algorithms outperform metabolite tests in predicting response of patients with inflammatory bowel disease to thiopurines. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 8, 143-150 (2010).

[73] Han L, et al. A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. Bioinformatics. 34, 985-993 (2018).

[74] Wang L, et al. Applying Machine Learning Models to Predict Medication Nonadherence in Crohn's Disease Maintenance Therapy. Patient preference and adherence. 14, 917-926 (2020).

[75] Taylor KM, et al. Genetic and Inflammatory Biomarkers Classify Small Intestine Inflammation in Asymptomatic First-degree Relatives of Patients With Crohn's Disease. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 18, 908-916.e913 (2020).

[76] Pal LR, Kundu K, Yin Y, Moult J. CAGI4 Crohn's exome challenge: Marker SNP versus exome variant models for assigning risk of Crohn disease. Hum Mutat. 38, 1225-1234 (2017).

[77] Doherty MK, et al. Fecal Microbiota Signatures Are Associated with Response to Ustekinumab Therapy among Crohn's Disease Patients. mBio. 9, (2018).

[78] Daneshjou R, et al. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Hum Mutat. 38, 1182-1192 (2017).

[79] Giollo M, et al. Crohn disease risk prediction-Best practices and pitfalls with exome data. Hum Mutat. 38, 1193-1200 (2017).

[80] Kang T, Ding W, Zhang L, Ziemek D, Zarringhalam K. A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data. BMC Bioinformatics. 18, 565 (2017).

[81] Waljee AK, et al. Predicting corticosteroid-free endoscopic remission with vedolizumab in ulcerative colitis. Aliment Pharmacol Ther. 47, 763-772 (2018).

[82] Tong Y, et al. Can natural language processing help differentiate inflammatory intestinal diseases in China? Models applying random forest and convolutional neural network approaches. BMC Med Inform Decis Mak. 20, 248 (2020).

[83] McDonnell M, et al. High incidence of glucocorticoid-induced hyperglycaemia in inflammatory bowel disease: metabolic and clinical predictors identified by machine learning. BMJ Open Gastroenterol. 7, (2020).

[84] Jiang P, Wu S, Luo Q, Zhao XM, Chen WH. Metagenomic Analysis of Common Intestinal Diseases Reveals Relationships among Microbial Signatures and Powers Multidisease Diagnostic Models. mSystems. 6, (2021).

[85] Waljee AK, et al. Machine Learning Algorithms for Objective Remission and Clinical Outcomes with Thiopurines. Journal of Crohn's & colitis. 11, 801-810 (2017).

[86] Isakov O, Dotan I, Ben-Shachar S. Machine Learning-Based Gene Prioritization Identifies Novel Candidate Risk Genes for Inflammatory Bowel Disease. Inflamm Bowel Dis. 23, 1516-1523 (2017).

[87] Wei Z, et al. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. Am J Hum Genet. 92, 1008-1012 (2013).

[88] Yu S, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. J Am Med Inform Assoc. 24, e143-e149 (2017).

[89] Sudlow C, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine. 12, e1001779 (2015).

[90] Hauben M, Reynolds R, Caubel P. Deconstructing the Pharmacovigilance Hype Cycle. Clin Ther. 40, 1981-1990.e1983 (2018).

[91] Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies. Int J Med Inform. 153, 104510 (2021).

[92] Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health &amp;amp; Care Informatics. 28, e100251 (2021).

[93] Olczak J, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. Acta Orthop. 92, 513-525 (2021).

[94] Norgeot B, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 26, 1320-1324 (2020).

[95] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ : British Medical Journal. 350, g7594 (2015).

[96] Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet (London, England). 393, 1577-1579 (2019).

[97] Luo W, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. Journal of medical Internet research. 18, e323-e323 (2016).

[98] Ibrahim H, et al. Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. Trials. 22, 11 (2021).

[99] Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 27, 2011-2015 (2020).

[100] Liu Y, Chen P-HC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users’ Guides to the Medical Literature. JAMA. 322, 1806-1816 (2019).

[101] Mongan J, Moy L, Kahn CE, Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell. 2, e200029-e200029 (2020).

[102] Finch H, Schneider MK. Classification Accuracy of Neural Networks vs. Discriminant Analysis, Logistic Regression, and Classification and Regression Trees. Methodology. 3, 47-57 (2007).

[103] Maroco J, et al. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res Notes. 4, 299 (2011).

[104] Olson RS, La Cava W, Orzechowski P, Urbanowicz RJ, Moore JH. PMLB: a large benchmark suite for machine learning evaluation and comparison. BioData Mining. 10, 36 (2017).

[105] Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:190602243. (2019).

[106] Eelbode T, Sinonquel P, Maes F, Bisschops R. Pitfalls in training and validation of deep learning systems. Best Practice & Research Clinical Gastroenterology. 52-53, 101712 (2021).

[107] Pannala R, et al. Artificial intelligence in gastrointestinal endoscopy. VideoGIE. 5, 598-613 (2020).

[108] Challen R, et al. Artificial intelligence, bias and clinical safety. BMJ Qual Saf. 28, 231-237 (2019).

[109] Yu KH, Kohane IS. Framing the challenges of artificial intelligence in medicine. BMJ Qual Saf. 28, 238-241 (2019).

**LEGENDS**

Figure 1 Flowchart documenting number of records found and reviewed at each stage.

Table 1 Summary of ML models chosen as most optimal for the clinical task, and the types of data used (ML models and data types sorted alphabetically).

Figure 2 Sample sizes used for each group of machine learning methods. BN = Bayes Network, DT = Decision Tree, NN = Neural Network, RF = random forest, SVM = support vector machine. Note that 10 outlier entries (sample sizes 20,368-7,400,000) have been excluded from the visualisation

Figure 3 Sunburst of machine learning methods, and the classification tasks used in conjunction with them.

Figure 4 Implementation of machine learning methods over time, incomplete data for 2021.