genes

MDPI

Review

Machine Learning Modeling from Omics Data as Prospective Tool for Improvement of Inflammatory Bowel Disease Diagnosis and Clinical Classifications

Biljana Stankovic *,†, Nikola Kotur †, Gordana Nikcevic, Vladimir Gasic, Branka Zukic and Sonja Pavlovic

Laboratory for Molecular Biomedicine, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, 11042 Belgrade, Serbia; nikola.kotur@imgge.bg.ac.rs (N.K.); gordnik@imgge.bg.ac.rs (G.N.); vlada.gasic@imgge.bg.ac.rs (V.G.); branka.zukic@imgge.bg.ac.rs (B.Z.); sonya@sezampro.rs (S.P.)

* Correspondence: biljana.stankovic@imgge.bg.ac.rs

† Authors contributed equally to this work.

Abstract: Research of inflammatory bowel disease (IBD) has identified numerous molecular players involved in the disease development. Even so, the understanding of IBD is incomplete, while disease treatment is still far from the precision medicine. Reliable diagnostic and prognostic biomarkers in IBD are limited which may reduce efficient therapeutic outcomes. High-throughput technologies and artificial intelligence emerged as powerful tools in search of unrevealed molecular patterns that could give important insights into IBD pathogenesis and help to address unmet clinical needs. Machine learning, a subtype of artificial intelligence, uses complex mathematical algorithms to learn from existing data in order to predict future outcomes. The scientific community has been increasingly employing machine learning for the prediction of IBD outcomes from comprehensive patient data-clinical records, genomic, transcriptomic, proteomic, metagenomic, and other IBD relevant omics data. This review aims to present fundamental principles behind machine learning modeling and its current application in IBD research with the focus on studies that explored genomic and transcriptomic data. We described different strategies used for dealing with omics data and outlined the best-performing methods. Before being translated into clinical settings, the developed machine learning models should be tested in independent prospective studies as well as randomized controlled trials.

Keywords: IBD; artificial intelligence; prediction modeling; genomics; transcriptomics

check for updates

Citation: Stankovic, B.; Kotur, N.; Nikcevic, G.; Gasic, V.; Zukic, B.; Pavlovic, S. Machine Learning Modeling from Omics Data as Prospective Tool for Improvement of Inflammatory Bowel Disease Diagnosis and Clinical Classifications. *Genes* **2021**, *12*, 1438. https://doi.org/10.3390/genes12091438

Academic Editors: Tamas Korcsmaros and Johanne Brooks

Received: 15 June 2021
Accepted: 14 September 2021
Published: 18 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Inflammatory bowel disease (IBD) is a complex disease, characterized as chronic, relapsing and remitting intestinal inflammation, with substantial heterogeneity among clinical phenotypes with regards to the age at diagnosis, severity of symptoms, response to therapy and long-term clinical outcomes [1–3]. It has traditionally been considered to comprise two major subtypes, Crohn's disease (CD) and ulcerative colitis (UC) [4]. This classification is mainly based on distinctive parameters primarily related to the location and behavior of the disease. CD can occur at various parts of the gastrointestinal (GI) tract-from mouth to anus, and it is patchy, transmural, and may have inflammatory (also called nonstricturing/nonpenetrating), stricturing or penetrating (fistulating) behavior [5]. UC is typically restricted to the colon and rectal mucosa, without fibrotic strictures [6,7]. In addition to CD and UC, there are patients whose disease characteristics cannot fit precisely into either of these two subtypes, which are described as 'IBD unclassified' (IBDU), and they are more common in children [8].

A number of interacting factors are accountable for the pathogenesis of IBD, of which genetic susceptibility, bacterial recognition and immune response of the host, microbiota and diet are among the most significant ones [9,10].

The search for IBD genetic determinants resulted in identification of more than 240 gene loci that have been associated with an increased risk of developing this disease [11–13]. In 2001 the frameshift variant of *NOD2* (nucleotide-binding oligomerization domain-containing protein 2) gene was identified as the first CD susceptibility genetic variant [14]. Currently, IBD is characterized as a polygenic disease, driven by multiple common genetic variants, of which *NOD2* variants have the highest effect size [12,15]. Also, it has been shown that rare monogenic variants contribute to the IBD risk, and to date, around 50 single genes are implicated in very-early-onset IBD [12,15,16].

Further studies revealed that the major perturbed molecular processes in IBD are associated with signaling pathways involved in innate and adaptive immune response, autophagy and intestinal epithelial barrier function and repair [17–21]. Although the etiology of IBD still remains undefined, the host-genome association with gut microbiome is in the focus of the current model of IBD pathogenesis. It is based on the concept of misdirected response of the host's immune system to intestinal microbial and immunogenic factors that involve the inflammation-associated mucosal injury [22]. It is believed that these are the key steps which promote disease severity, relapse and also its progression to neoplastic transformation [12,23,24].

Currently, there is no cure for IBD, and in a significant number of cases, applied therapies are found to be ineffective or lead to a poor/inadequate response [25–27]. In addition, it is often not possible to establish an accurate diagnosis of IBD since it depends on a combination of numerous clinical data, including complex image assessments, whose interpretation is inherently subjective [28]. Altogether, untimely and inaccurate diagnosis has a great impact on the course of the disease, which usually leads to complications and thus represents a serious obstacle to achieving and maintaining remission of the disease, which is the main goal of the IBD treatment [29]. Diagnosis, classification, prognosis and therapy of IBD still require the detection of accurate and reliable biomarkers and their translation into clinical practice, with the aim to significantly improve outcomes in patients with IBD.

Regarding molecular classification of disease subtypes, it has been shown that most of detected IBD loci confer risk to both CD and UC, typically with distinct effect sizes in each disorder; whereas the minor number of loci is unique to each subtype [15,30]. In addition to the latter, there are examples, such as variants at the *NOD2* and protein tyrosine phosphatase nonreceptor type 22 (*PTPN22*) loci, which have been found to be risk factors for CD, while for UC they have been shown to be variants with a protective effect [30]. These results provided evidence for fundamental etiological differences between the two IBD subtypes [21]. In addition, important connections have been found between molecular and clinical phenotypes, such as the ones related to disease location (associations of *NOD2*, *MHC* and *MST1* 3p21 variants with ileal vs. colonic CD) and to disease behavior (associations of *NOD2*, *MHC* and *MST1* 3p21 variants with deep ileal ulcers; *NOD2* variants with fibrostenotic or stricturing ileal CD; miR-215 expression with penetrating CD and differential DNA methylation with inflammation in CD) [21].

Cumulative evidence suggests that classifying IBD as CD or UC might be oversimplified and that particular disease phenotypes could also be considered as genetically distinct entities [11]. Namely, results of the large genotype-subphenotype study performed on the Immunochip array showed that colonic CD was genetically intermediate between ileal CD and UC; i.e., predictive models based on genetic risk scores identified that ileal CD and colonic CD are as different from each other as they are from UC [11]. This finding corroborated the results of several earlier studies, which have found that *NOD2* gene variants are associated with ileal Crohn's disease, thus delineating it from colonic CD, with a shorter time for the onset of stenosing disease and need for surgery [12]. Important additional evidence is that differential microRNA (miRNA) expression was detected between these entities [31]. In this sense, it has been suggested recently that ileal and colonic CD could potentially be regarded as separate diseases and that consideration should be given to a new classification for CD, which splits it into ileum dominant (isolated ileal and ileocolonic) and isolated colonic disease. This may allow for a more optimized approach to clinical care and scientific research for CD [32].

The integration of data from expression (mRNA, miRNA and protein) and epigenetic (DNA methylation and histone modifications) examinations is progressively more present in IBD studies, which significantly contribute to the improvement of IBD classification, reduce misdiagnoses and assist clinical decisions regarding the choice of adequate therapies [21,33–36]. The emergence of new high-throughput technologies enabled the usage of genomic, transcriptomic, epigenomic, proteomic, metabolomic, metagenomic, in general—omics data, for the purpose of achieving the goals of precision medicine. The clinical diagnostics based on omics analysis using next generation sequencing are applicable, especially in the fields of inherited disease diagnosis and oncology. Several omics-based tests have been FDA (Food and Drug Administration) approved for the clinical application in the USA, as well as certified with CE (Conformitè Europëenne) marking for clinicians in Europe. Most of the tests comprise genome analysis, but there are also tests using RNA whole-transcriptome sequencing (https://www.clinicalomics.com, accessed on 25 November 2020).

The analysis of omics big data demands the usage of powerful bioinformatics tools and application of advanced statistics such as artificial intelligence (AI). Machine learning (ML), a subset of AI, is the most promising tool nowadays in search of new clinically relevant patterns and reliable predictive markers of complex diseases [28,37]. The underlying genetic predisposition to IBD has not been completely revealed employing only the candidate gene approach or genome wide association studies (GWAS). For that reason, there is great interest in the application of AI in IBD research with the goal to improve: patient identification, differential diagnosis, disease risk prediction and clinical outcomes and classification of disease subtypes, as well as identification of disease biomarkers that could be targeted for advancing therapeutic management (Figure 1) [1,28,38,39].

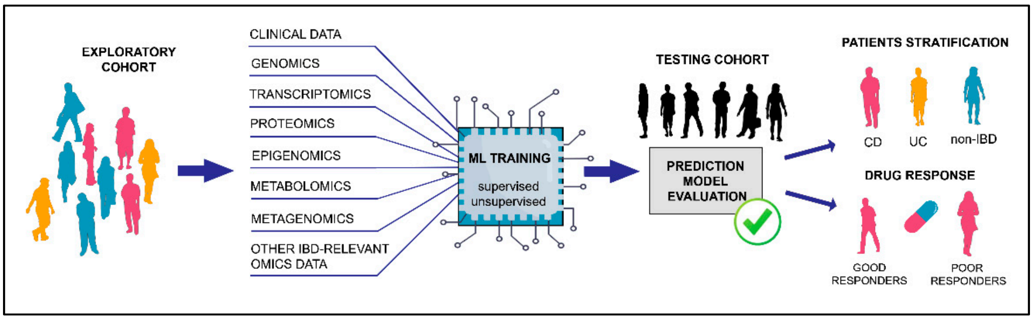

Figure 1. Machine learning using omics data for prediction of clinically relevant IBD outcomes. Omics data from patients with known clinical outcomes (exploratory cohort) can be used as input data in machine learning algorithms during the prediction model training. Performance of the designed model is further assessed on an independent group with unknown outcomes (testing cohort). Machine learning models that have high prediction performance on the testing cohort are well fitted and could be employed for future improved patients' diagnosis, classification, prognosis and prediction of drug response. ML—machine learning, CD—Crohn's disease, UC—ulcerative colitis, and IBD—inflammatory bowel disease.

Besides omics, different clinical measurements used for IBD diagnosis and tracking of the disease status, such as fecal calprotectin, blood parameters, serum C-reactive protein, endoscopic and/or medical imaging, possess a large potential that could be exploited in machine learning modeling. A number of studies analyzed usage of clinically valuable traits in IBD diagnostic, prognostic and therapeutic outcome predictions [40–45]. For instance, it has been demonstrated that machine learning algorithms employing laboratory and age data outperformed drug metabolic measurements in predicting the response of IBD patients to thiopurines [41]. A promising machine learning utility in IBD is expected for artificial-intelligence-assisted medical images analysis, which is a more objective and computable technique that can automate and improve intrinsically subjective endoscopic evaluations [46,47]. This could help less experienced endoscopists and reduce interobserver variability. Essentially, the real potential of personalized medicine lies in integration of clinical and multiomics data.

This review discusses machine learning application in IBD with the focus on studies that explored genomic and transcriptomic data. First, the basic concepts of machine learning and the foundation of the most used algorithms were explained. Then, we evaluated the representative studies that employed machine learning on genomic and transcriptomic datasets for predicting IBD clinical outcomes or identification of novel risk genes. Finally, we argued the future perspectives of AI in IBD research and prerequisites for its successful translation into clinical practice.

2. Machine Learning Approaches

Machine learning is an important area of AI that provides a machine with an ability to learn from experience or find patterns in the data without being explicitly programmed. ML employs self-learning algorithms (set of rules) to solve classification and regression problems (supervised learning) or to find hidden patterns (unsupervised learning) in data. Short descriptions of the common terms related to machine learning used in this review are summarized in Table 1.

Table 1. Glossary of common terms in machine learning.

Instance	An entity (human subject in healthcare applications) which features are used as inputs for prediction modeling.
Feature	An explanatory variable, such as genetic variant, gene expression, etc. Features are used as input data for prediction.
Machine learning algorithm	Procedure that is run on data to create a machine learning model. It is a set of mathematical optimization functions that minimizes the error of the model function.
Iterations	Machine learning algorithm's parameters are updated number of times until model reaches desired performance
Classification	Supervised learning technique used to predict a discrete class or category of an instance (disease or healthy subject, good or poor drug responder, etc.).
Regression	Supervised learning technique in which the predicted variable is continuous.
Model fitting	Measure of how well a machine learning model generalizes to data not used for model training.
Penalized regression method	A method used to reduce overfitting of a model. The penalty causes the regression coefficients of less contributive variables to shrink toward zero therefore reducing the number of variables in the model.
Sparse model	A predictive model that includes only the most informative features.
Clustering	Unsupervised learning technique that groups instances by their similarity. The groups are called clusters.
Black box model	Model that is built on complex functions that are not easily interpreted (such as neural networks). Input and output are clear, but the process between is not explainable.
Effect size	A biological measure of the difference or relationship between variables. An OR << 1 or OR >> 1 is indicative of a large effect size.
AUC value	Evaluation metric of a model that ranges from 0.5 (poor classifier) to 1 (perfect classifier).

Supervised machine learning algorithms are used to solve classification problems or predict response value (regression) based on historical, example data. Example data (the training set) contains labeled instances (usually human subjects in healthcare applications), which means that both input (features or explanatory variables) and output, which is the phenotype of interest (the response variable), are known. Using ML, researchers can obtain an approximate function or a model that successfully differentiates between classes or predicts the numerical value of the response variable. In the designed model, the underlying codependency between the input and output variables often is mathematically complex and not easily interpreted. Successful models should give accurate predictions when applied to novel instances which have not been used for model training. In practice, different types of algorithms are used to learn the model function [48]. The most commonly used classification and/or regression algorithms in IBD research are linear algorithms, support vector machines (SVM), k-nearest neighbors, decision trees, Bayesian algorithms and artificial neural networks [37,49–51] (Figure 2).

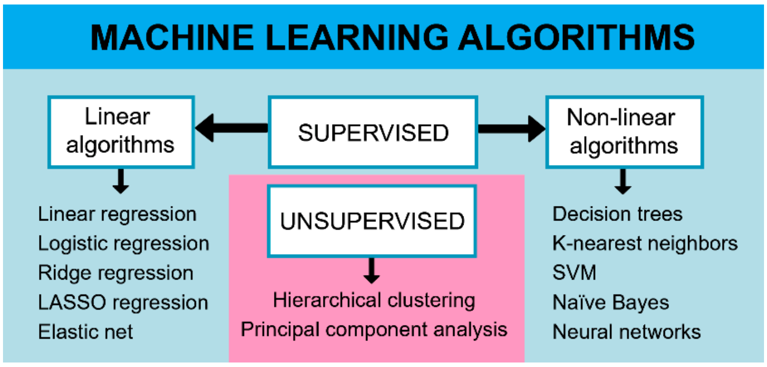

Figure 2. Classification of machine learning algorithms used in IBD research; LASSO—least absolute shrinkage and selection operator; SVM—support vector machines.

Unsupervised learning deals with unlabeled data, and the aim is to group instances according to similarity and to find structures within the data. This approach is particularly useful when multiple input variables are included because researchers are unable to visualize and find patterns in hyper-dimensional space. Also, these methodologies are also used for anomaly detection and dimensionality reduction. In life science and IBD research, unsupervised learning algorithms such as hierarchical clustering and the principal component analysis are commonly used [52,53] (Figure 2).

2.1. Fitting the Model

In supervised learning, a model represents a function that captures codependency between input variables and the known outputs in order to predict the future unknown outputs of novel instances. Usually, several models are developed in an attempt to obtain a well-fitted model using different algorithms. Algorithms often use iterative protocols to fit a model function to the data. For example, the gradient descent protocol tunes parameters of the model function in small steps toward (local) minimum of the error function, which represents the measure of deviation of the model function from the example data [48]. A model should ideally capture a general trend within the data but not random noise and erroneous measurements. If a model is trained too long (too many iterations) or if it is too complex, the model captures irrelevant details of the example data and does not generalize well when exposed to new data. This leads to overfitting of the model. This problem could be mitigated by limiting the number of iterations or by penalizing the complexity of the model. The commonly used method to assess the (over)fitting of a model is based on splitting the data into training, validation (or development) and testing sets. The training set (exemplary data) is used to fit a model function. The validation set is used to tune the parameters toward a less complex model that would generalize better. Finally, the testing set is used to make the final assessment of predictive performance of the model. Resampling-based techniques such as bootstrap resampling and cross-validation provide an opportunity to use a single dataset for both training and validations. Here, a model is trained of the majority of instances, and only a small part is randomly chosen and set aside for validation. This procedure is usually iteratively performed to obtain robust assessment of the error function. Resampling-based techniques are useful when only a small number of instances is available [54]. Another problem with predictive modeling occurs when neither the training nor validation set is well fitted by a model function, which is called underfitting. This problem is easily noticed by examining predictive performance of the model on the training set. In case of underfitting, other model functions and algorithms should be tried in order to develop a well-fitted model.

Classification model is evaluated according to its accuracy, which is the percentage of correct predictions. Another popular metric used to assess and visualize the predictive performance of classification models is area under the receiver operating characteristic (ROC) curve or AUC. The curve captures codependency between true positive rate and false positive rate of the model. A similar metric, area under the precision-recall curve or AUPRC, captures codependency between precision (or positive predictive value) and recall (or sensitivity). Both AUC and AUPRC provide a single-value metric to evaluate predictive performance of a classification model, which normally ranges between 0.5 (poor predictive performance) and 1 (perfect classifier) [55].

Different algorithms are used to fit a model function to the data. The choice of an algorithm depends on the problem (classification, regression, clustering, etc.) and the dataset (number of features, number of instances, codependency between features, etc.).

2.2. Linear Algorithms

Linear algorithms assume linear dependency between input variable(s) and the output. Assumed linear dependency provides a more straightforward interpretation, because the contribution, both the sign and the effect-size, of each input variable to the model is known. Another advantage of linear models is that they are computationally less expensive to develop. On the other hand, linear models might perform poorly if dependency between input variables and the output is not linear, which would lead to underfitting. In the case of IBD research, there is still no strong evidence that nonlinear models outperform linear ones [37,51,56].

Linear regression is one of the most understood machine learning algorithms used to predict continuous output variables. The algorithm develops a linear model function that best fits the data around a straight line (or a hyperplane). Logistic regression, on the other hand, is not a regression instrument; instead, it is used for classification into distinct categories. Here, a linear function is transformed into a sigmoid-shaped line (the logistic function) which best differentiate between categories. Both linear regression and logistic regression suffer from overfitting and challenging interpretation if multiple input variables are included into the model. To deal with these issues, the modified linear algorithms based on regularization, such as the ridge regression (L2), least absolute shrinkage and selection operator (LASSO/L1) and elastic net (both L1 and L2) can be employed (Table 1). These algorithms penalize the complexity of the model function by shrinking coefficients coupled with input variables toward zero. The coefficients of less predictive input variables can shrink exactly to zero (often encountered with L1 regularization), which would effectively exclude those variables from the model. Another regularization strategy is to shrink all coefficients more evenly (L2 regularization), which effectively deals with codependent features [57].

2.3. Nonlinear Algorithms

Nonlinear algorithms do not assume the shape of model function and provide more flexibility and, therefore, better opportunity to develop a well-fitted model. However, they are often more computationally expensive to develop, more prone to overfitting and harder to interpret than the linear models. In addition, these algorithms usually require many instances to provide optimal results [48]. Commonly used nonlinear algorithms include decision trees, k-nearest neighbors, support vector machines, naïve Bayes and neural networks (Table 2). Deep learning is a very popular extension of neural network algorithms which employs multiple hidden layers of interconnected artificial neurons stacked between the input and the output [58].

Table 2. Classification and regression machine learning algorithms employed in IBD research.

Algorithm	Principle	Usage	Pros and Cons
Logistic regression	Linear model transformed into sigmoid function used as a binary classifier	Classification	Fast to develop; easily interpretable; limited by strong assumptions; prone to overfitting
Linear regression	Classical linear model that employs linear codependency for prediction	Regression (can also be used for classification)	Fast to develop; easily interpretable; limited by strong assumptions; prone to overfitting
Ridge regression	Linear model with L2 regularization	Classification and regression	Linear model with enhanced interpretability and reduced overfitting
LASSO	Linear model with L1 regularization	Classification and regression	Linear model with enhanced interpretability and reduced overfitting
Elastic net	Linear model with both L1 and L2 regularization	Classification and regression	Linear model with enhanced interpretability and reduced overfitting
Decision trees	Prediction based on a tree-like model. Nodes are splitting points of a dataset based on most informative features; leaves are output values.	Classification and regression	Prone to overfitting but can be improved with ensemble methods; interpretable outputs
Random forest	An ensemble method (modified bootstrap aggregation) applied to decision trees. It grows multiple decision trees; output is the average prediction of individual trees.	Classification and regression	High prediction performance; deals with overfitting; requires a large dataset for optimal learning.
Gradient boosted trees (GBT)	An ensemble method (gradient boosting) applied to decision trees	Classification and regression	High prediction performance; hard-to-tune parameters of the algorithm
K nearest neighbors (KNN)	Predicts an output taking into account (k) most similar instances (nearest neighbors)	Classification and regression	Requires a lot of memory to store all the instances; cannot deal with a large number of input variables.
Support vector machines (SVM) classifier	Maximizes margin (decision boundary) between different classes supported by instances that lie near the margin (support vectors)	Classification	Works well with high number of input variables; flexible (allow curved margin by using nonlinear kernels); computationally expensive; limited interpretability
Naïve Bayes	Employs Bayesian posterior probability theorem but assume nondependency between features given the output	Classification	Fast to develop; suitable for large datasets and for making real time predictions; limited by strong assumptions; requires feature selection and transformation
Neural networks	Network of interconnected units resembling the nervous system which renders input information to produce an output.	Classification and regression	High performance; limited interpretability; requires very large dataset; computationally expensive

LASSO—least absolute shrinkage and selection operator.

Classification and regression models often suffer from a poor prediction performance either because of overfitting or underfitting the training data. The so-called "ensemble methods" can address these issues by combining predictions from multiple (usually weak) models, which delivers better predictive performance [48]. Among the commonly used ensemble methods are random forest and gradient boosted trees (GBT), both enabling higher prediction performance of decision tree algorithms [37,49] (Table 2).

2.4. Clustering Algorithms

Clustering algorithms group instances based on similarities or distance in feature space. As it is an unsupervised ML approach, the number of clusters is not predetermined. Hierarchical clustering iteratively groups instances into larger clusters until being merged into a single cluster. The clustering process is captured in a tree-like structure which expectantly reflects the underlying organization of the data. Bayesian hierarchical clustering is also a bottom-up approach in which statistical testing guides grouping of the clusters [59].

3. Machine Learning in IBD Research

ML methods are currently the best tools for dealing with complex omics data in IBD prediction. The main issue in standard association genotype–phenotype studies using omics data is the large number of multiple comparisons that require rigorous statistical methods for avoiding false positive results. Because of that, many potential causal variants with usually small effect sizes are being neglected. By contrast, ML approach is more flexible in recognizing disease patterns regardless of the statistical level of the associated variants [60]. Only a small percent of IBD heritability is currently explained by identified risk loci [12]. The genetic architecture of IBD is polygenic, with both rare and common variants contributing to disease risk. The large effect sizes have single causal variants (*IL10* and *XIAP*), followed by high-risk variants (odds ratio [OR] > 2) (*NOD2* fs1007insC, *CARD9* c.1434+1G>C, *HLA-DRB1*, etc.), then medium-risk variants (OR 1.2–2) (*NOD2* Asn289Ser, *IL23R* Val362Ile, etc.), to the common disease susceptibility loci, which have small effect sizes (OR < 1.2), collectively accounting for only a fraction of variance in disease liability [12]. ML has the ability to select predictor genes with small contributions and to capture effects of epistasis (interactions between genes) which is very important for complex diseases. Thus, ML may further resolve genetics of IBD and indicate new relevant pathways, undiscovered before by standard statistics.

There has been an increased interest in recent years in using AI to explore omics data for IBD risk prediction and classification [37,49,51,61]. The designed ML models had variable prediction performance, with AUC ranging from 0.7 to 0.95, depending on the used dataset and applied method (Table 3). The most frequently employed ML methods included penalized regression models, random forest, support vector machines, Bayesian approach and neural networks (Table 3). Even though ML models are often "black boxes", they could be used for identifying potentially causal molecular patterns of IBD by evaluating the most significant genes/features selected during the process of model training [52,53,62–64].

ML algorithms are data hungry and need large sample sizes to obtain the best performance. The whole process of ML assumes iterative steps of model training with validation followed by model testing on the independent dataset. The common practice in the majority of ML studies performed on IBD is to prefilter variants for subsequent modeling [51,65,66]. In this way, high computational costs caused by high-dimensional input data can be reduced while the overfitting problem is avoided. Even so, the best strategy for detecting all causal variants of complex diseases might be the application of sparse penalized models on the whole set of genotyped variants [60].

Taking into account the price of genome and transcriptome analysis, prediction modeling using omics data on large cohorts could be expensive. However, as the price of the NGS and other high-throughput techniques decreases over time, more frequent application of ML using omics data in IBD risk predictions is expected. Still, beside the greater availability of the high-throughput techniques, achieving good predictive results is often limited due to widespread presence of confounding effects, relatively low prevalence of IBD and high heterogeneity of the disease phenotypes [63]. These issues often limit the analyzed sample size or make the dataset less uniform. Large IBD consortiums having collected and analyzed tens of thousands of samples along with promoting open-access data are an extremely valuable source of omics datasets which could be extensively explored in IBD prediction modeling.

Table 3. Studies that explored machine learning for designing IBD prediction models using genomic and transcriptomic data.

First Author and Year [ref]	Machine Learning Algorithm	Predictors/Prediction	Performance	Tested on Independent Cohort	Subjects
Chen 2017 [65]	Bayesian mixture approach	GWAS or Immunochip SNPs data/IBD risk score	CD AUC: 0.75, UC AUC: 0.70	yes	The IIBDGC) cohort—over 68,000 IBD patients and 29,000 healthy controls (4:5 ratio for training and testing, respectively)
Wei 2013 [66]	L1 penalized logistic regression, SVM, gradient boosted trees	Immunochip SNPs data/CD and UC distinction from healthy controls	CD AUC 0.86, UC AUC 0.83	yes	The IIBDGC cohort—~17,000 CD, ~13,000 UC, and ~22,000 controls (randomly divided into 3 folds of equal size for preselection, training and testing, respectively)
Romagnoni 2019 [37]	Logistic regression, gradient boosted trees, neural network and ensemble method	Immunochip SNPs data/probability of CD	AUC 0.8	yes	The IIBDGC cohort—train dataset (34,634 samples), test dataset (17,317 samples)
Pal 2017 [51]	Naïve Bayes	Exome data/CD status	AUC 0.81	yes	Training set: 64 CD and 47 controls (CAGI4); Testing set: 51 CD and 15 controls (CAGI3)
Raimondi 2020 [63]	Neural network	Whole exomes/to distinguish between CD and healthy controls	AUC 0.74–0.83 AUPRC 0.81–0.93	yes	CAGI2, CAGI3, CAGI4 datasets (training and testing)
Wang 2019 [64]	SVM	Whole exomes/to distinguish between CD and healthy controls	AUC 0.7–0.75 AUPRC 0.73–0.80	yes	CAGI4 (training set), CAGI3 (testing set)
Isakov 2017 [49]	Random forest, SVM with polynomial kernel, extreme gradient boosting, elastic net and ensemble method	Data from 2050 genes annotated by the expression (array and RNAseq) and pathway information (categorical terms)/IBD-risk gene prioritization	AUC 0.775–0.829	yes	Intestinal biopsies of 180 CD, 149 UC, 94 colorectal neoplasms, 90 normal tissue (75:25 ratio for training and testing set, respectively)
Cushing 2018 [52]	Unsupervised hierarchical clustering, random forest	Whole transcriptome/identification of markers that could predict postoperative disease activity	92–93% of correct estimates in random forest	no	24 anti-TNFα-naïve patients, 30 anti-TNFα-exposed
Khorasani 2020 [53]	Feature selection algorithm (based on dimension reduction) combined with SVM classifier	Wide expression array data/UC and healthy subjects classification	Active UC AUPRC 1, Inactive UC AUPRC 0.68	yes	Training set: 39 UC samples (active and inactive) and 38 controls; testing set: 97 UC samples (active and inactive) and 22 controls

Table 3. *Cont.*

First Author and Year [ref]	Machine Learning Algorithm	Predictors/Prediction	Performance	Tested on Independent Cohort	Subjects
Yuan 2017 [62]	Feature selection (minimum redundancy maximum relevance and incremental feature selection), SVM-based algorithm (sequential minimal optimization)	Wide expression array data from PBMC samples/CD, UC and normal subject discrimination and candidate gene selection	Accuracy 0.94	no	59 Crohn's disease, 26 ulcerative colitis, and 42 normal samples
Hubenthal 2015 [67]	Penalized SVM, random forest	miRNAs in whole-blood samples/IBD and control subject distinction	AUC 0.75-1.0	no	40 CD, 36 UC, 38 healthy controls and other inflammation controls (24 chronic obstructive pulmonary disease, 23 multiple sclerosis, 38 pancreatitis and 45 sarcoidosis cases)
Zarringhalam 2014 [68]	Differential expression profile was used to infer upstream regulators using Bayesian approach; posterior probabilities of regulators' activities were then used in a regularized regression framework to predict outcome	Genome wide expression data/response to infliximab in UC	Accuracy 0.79	yes	Training set: 22 active UC patients (12 responders and 10 nonresponders); Testing set: 24 active UC patients (8 responders and 16 nonresponders)
Li 2020 [50]	Random forest, neural network	RNAseq and microarray expression data/identification of susceptibility genes and establishing predictive model of UC	AUC 0.95; AUPRC 0.97	yes	Training set: 206 UC, 20 normal; Testing set: 53 UC and 21 normal
Martin 2019 [69]	Hierarchical clustering, principal component analysis	Single-cell RNA sequencing data/cell type classification in inflamed and uninflamed tissues	Inflamed tissue (r = 0.96) Uninflamed tissue (r = 0.93) *	no	11 ileal CD patients; samples taken from inflamed and uninflamed tissues

GWAS—genome-wide association study, IBD—inflammatory bowel disease, CD—Crohn's disease, UC—ulcerative colitis, SVM—support vector machine, AUC—area under the receiver operating curve, AUPRC—area under the precision-recall curve, IIBDGC—The International Inflammatory Bowel Disease Genetics Consortium, *—correlation of cell type frequencies between hieratical clustering analysis applied to RNA profile of a cell and cytometry results referring to that cell.

3.1. Machine Learning Using Genomic Data

Currently, the largest available IBD genomic dataset has been provided by the International IBD Consortium (IIBDC). The IIBDC dataset was used by Chen and colleagues to predict IBD risk scores [65]. This dataset consists of GWAS imputed and Immunochip genotyped SNPs from over 68,000 IBD patients and 29,000 healthy controls that enabled discovery of more than 200 risk IBD loci [17,30]. Immunochip is a custom Illumina assay comprising 196,524 SNPs and small indels selected primarily based on GWAS analysis of 12 autoimmune and inflammatory diseases [30]. In their analysis, Chen et al. varied the methods for estimating IBD risk score, sample size and type of data used for prediction (GWAS or Immunochip). Their study pointed to Bayesian hierarchical clustering as the best performance algorithm. In addition, they showed that Immunochip data had similar prediction performance as GWAS, largely due to the guidance of the initial GWAS for the Immunochip marker selection. This study indicated that the power of genomic CD and UC prediction was mainly due to strongly associated SNPs with considerable effect sizes. Additional SNPs tagged by GWAS arrays and rare variants found on the Immunochip contributed little to prediction accuracy. Other studies as well came to similar conclusions. The inclusion of not only a significant but broader set of variants, as well as the addition of rare alleles in IBD-established genes, did not improve disease risk prediction performance [37,51]. These results were in contrast with the expected potential of ML to reveal genetic variants carrying marginal IBD risk effects.

Wei and coworkers also used the IIBCD dataset [66]. The study yielded an IBD risk prediction model with high performance (AUC 0.860) using the penalized logistic regression method. The authors applied a two-step feature selection strategy: first, features (genetic variants) were filtered after single-association tests by less stringent association significance cutoff ($<10^{-4}$) and taking into account the frequency of the minor allele (>0.01), and then, LASSO (L1) penalization was performed on the remaining variants. Given the size of the dataset, the LASSO penalization approach was chosen because it requires only one parameter to be tuned during the process of optimizations, which decreased the high computational cost of the analysis.

Another study that exploited the IIBDC Immunochip data was conducted by Romagnoni and colleagues [37]. The authors aimed to make predictions of CD probability employing a set of ML methods: penalized logistic regression, gradient boosted trees and artificial neural networks. All ML methods showed AUC values in similar ranges. The slightly increased performance was accomplished using the ensemble method that combines logistic regression, gradient boosted trees and neural network classifiers, indicating that different models can be seen as partially complementary. This study pointed to several important conclusions—that quality control, imputing methods for missing genotypes and coding strategies for input data can affect the performance of the model, inducing artificial increase in the AUC scores [37].

One great example of the community experiment is the critical assessment of genome interpretation (CAGI), which aims to advance ML methods for genotype–phenotype prediction. CAGI provides a platform for assessing training and testing datasets (https://genomeinterpretation.org, accessed on 2 April 2021) which participants can use to make blind predictions. Since 2010, CAGI has presented dozens of datasets, so-called "challenges". During each challenge, the CAGI organizers release unpublished data and formulate a specific task related to it. After the closure, organizers evaluate performances of submitted predictions, and a conference is organized to discuss results and emerging ideas. This common task framework led to significant insights into ML-related problems.

In the years 2011 (CAGI2), 2013 (CAGI3) and 2016 (CAGI4), researchers tried to distinguish between CD and healthy controls based on whole exome data. CAGI 2, 3 and 4 datasets' sizes were not large, counting 56, 66 and 111 exomes, respectively. The work on these challenges stressed the critical points of ML application in genomics—discovering hidden biases in datasets, finding the best strategies to reduce data dimensionality and dealing with limited sample size [70,71].

One successful submission in CAGI4 was performed by Pal and coworkers [51]. They reduced the number of predictors by filtering exome data, including only genomic regions previously associated with CD [30,72]. The authors tested four ML algorithms—logistic regression, random forest, naïve Bayes and a neural network—and varied the number of genetic loci incorporated into the model (90 vs. 138). The best performance was achieved with naïve Bayes. A higher number of included loci improved prediction accuracy.

In a recent study by Raimondi et al., the authors designed a novel neural network approach model, CDkoma, to classify CD from healthy controls using CAGI 2, 3 and 4 editions exome data [63]. Initially, the established CD associations were used for selection of predictor genes using PhenoPedia [73]. The authors further dealt with high-dimensionality of the dataset applying efficient encoding strategy. Before entering the model neural nodes, the genetic variants were firstly aggregated at the gene level by counting how many times each type of variant occurs in each gene. This minimized complexity of the training data and the issue of overfitting, making this approach particularly suitable for the small size datasets. Interestingly, this study attempted to "open the neural network black box" and allow a biological interpretation derived from the ML model, even though the neural networks are known to be one of the most difficult ML to interpret.

Similar to the Raimondi study, Wang et al. applied gene-level encoding strategy [64]. For each gene in the set, the gene function score was computed on the basis of predicted functional effects of all its variants. This scoring system was far better compared to the one that calculated the total number of risk variants per gene [63,64]. Wang analyzed the performance of SVM model with leave-one-out cross-validation on CAGI4 as CD-train and CAGI3 as CD-test dataset. Selecting genes in the process of computational feature selection without any previous knowledge of CD biology gave better results than choosing predetermined GWAS genes (AUC 0.75 vs. 0.70, respectively). This suggests that functional effects of variants are more likely to highlight causative signals rather than association signals. Only a few genes appeared both in the feature selection and experimentally derived (GWAS) sets, implying that computational feature selection could identify previously unknown CD-related genes and could be the best choice for analyzing complex diseases where suspect genes are not established or GWAS studies data are not available.

It has been estimated that the prediction of IBD and particularly CD, given its high heritability, should be able to achieve a maximum AUC between 0.96 and 0.98 by genomic profiling (assuming that all risk loci and their effect sizes are known) [74–76]. Even though this number seems to be promising, it should be noted that the low prevalence of the IBD limits the utility of genetic prediction. If the prevalence of the disease is low, for instance 1% with theoretic AUC of 0.98, only 12% of individuals who test positive develop the disease [74]. However, IBD risk prediction is hardly ever required for testing in the general population. Subjects who have family history of IBD or are at higher risk of having unresolved gastrointestinal symptoms or undetermined CD or UC diagnosis represent a distinct population in which the incidence of IBD is much higher. Moreover, genetic prediction may be used in existing patients to classify them in disease subphenotypes, to infer course of the disease and treatment response [77]. Therefore, the clinical utility of these models could be more important for higher risk groups and diagnosed patients than for the general population.

3.2. *Machine Learning Using Transcriptomic Data*

Apart from genomics, other fields of omics, such as transcriptomics, have been explored in IBD risk predictions. The search for reliable IBD biomarkers outside of purely genetic studies is emphasized by the fact that all IBD-associated genetic factors identified so far can explain only 20–25% of described cases, a small fraction of IBD variance and variability within subphenotypes [1,78].

Isakov et al. developed ML-based gene prioritization method to differentiate IBD-risk genes from non-IBD genes [49]. The supervised method was generated to produce two outputs—positive, if the gene had previous GWAS established IBD associations, and negative, if the gene had no association with IBD. Each gene was characterized with gene expression data and gene annotation features, which were used to construct the prediction model. Using the selected features, Isakov et al. trained four different ML models to produce gene risk scores: random forest, SVM, gradient boosting and elastic net. The range of each risk score was from 0 to 1, which corresponded with the level of confidence in which a gene is considered to be an IBD risk gene. The method was then used to assign the risk scores to the comprehensive list of 16,390 genes. The model has selected 347 genes with high prediction scores for IBD risk; among which, 163 were already known IBD genes, 117 genes had at least one publication associated with IBD, and for the residual 67, no existing research was found. This is a good example of how vast data existing in the public domain may be used to discover novel IBD-associated genes.

A recent study by Smith et al. used comprehensive transcriptomic data from the recount2 [79] database to address different predictive problems related to phenotype classification [56]. The recount2 database contains the analysis-ready RNAseq count data from genotype tissue expression (GTEx) project, the cancer genome atlas (TCGA) and the sequence read archive (SRA). The aim of the Smith study was to test ML in predicting numerous binary and multiclass phenotype outcomes; among which, two were related to IBD. Particularly, they used colon tissue transcriptomic data to classify three types of CD-B1 (inflammatory), B2 (stricturing) or B3 (penetrating/fistulating) behavior as well as to predict etrolizumab response in UC patients. The study analyzed the impact of normalization techniques, different sizes gene sets and ML techniques such as logistic regression, random forest and k-nearest neighbors. It was demonstrated that multivariate predictors outperformed predictors based on the single gene and that larger gene sets were more informative compared to smaller ones. In addition, L2-regularized regression applied to the centered log-ratio transform of transcript abundances was shown to be the best choice for predictive analyses using transcriptomic data.

Unsupervised ML methods could be utilized to categorize patients without any previous assumptions and obtain potently better classifications than existing ones. For instance, hierarchical clustering has been used to assess the classification of operated TNF-naïve CD patients using transcriptome signature in ileum mucosa [52]. It has been shown that patients with a Rutgeerts score of i0 (measure of disease activity at the follow-up colonoscopy) largely segregate together and are independent of patients with scores i1-4. Moreover, i0 vs. i1-4 segregation was better than between i0 and i1 vs. all other scores, even though i0 and i1 are usually considered to be signatures of clinical endoscopic remission. When this differential classification was further analyzed in a random forest model, a set of 30 transcripts was selected as the most influential in the model. Transcripts involved in the regularization of Bcl-2 and Bax-mediated apoptosis, a cell process discussed before as potentially significant for ileal CD subtype categorization [80], were identified among the significant predictors of postoperative remission. The prediction model demonstrated high accuracy: 92–93% of estimates in random forest were correct.

Studies performed on IBD suggest that genetic contribution is weaker in UC compared to CD [81,82]. Thus, using gene expression data for diagnostic, prognostic and classification purposes might be more appropriate for UC than using genomic data. The recent study by Khorasani et al. [53] designed a predictive model to discriminate UC and healthy controls using colonic transcriptome data. Datasets were selected from different studies to reduce the effect of technical conditions, and the training and validation sets were independent. Prior to ML modeling, the authors applied a novel feature selection method in order to reduce input data dimensionality to 32 genes, which were further used in an SVM classifier. The final model was able to classify active and inactive UC from healthy donors with average precision of 1 and 0.62, respectively. From the selected 32 genes, most did not have a direct link with IBD phenotype, and some were related to IBD-associated comorbidities, such as altered blood pressure, cholesterol level or colorectal cancer. One more example where ML modeling with its hypothesis-free approach could be a useful tool to identify novel risk genes whose role in IBD would be investigated further on.

There are many ways to perform the selection of genes that contribute the most to disease classification. Yuan and colleagues searched for the genes expressed in IBD blood samples which could be used to classify CD and UC from non-IBD subjects [62]. They applied a two-step feature selection method. First, the genes were ranked according to their relevance to the sample class label and their mutual redundancy. In the second step, incremental genes/feature selection was used in SVM model with a tenfold cross-validation to obtain most optimal combination of genes for discrimination of UC, CD and healthy samples. This approach yielded a set of 21 genes that could predict a diagnosis with high accuracy (93.7%). The obtained set of genes was extended with an additional 20 genes after evaluating the interaction network of proteins coded by these genes. Gene ontology pathways enriched with identified genes have been recognized before as important to IBD, such as the T-cell receptor signaling pathway, cell activation, and apoptosis.

Independent validation is a critical step in the development of any biomarker, assay or prediction model, in which how well they perform on the unseen data can be tested. The successful validation of IBD biomarkers was demonstrated in Biasci and colleagues' prospective study [83]. In the previous work of the authors, unsupervised clustering of CD8 T-cells transcriptome data separated IBD patients into two distinct subgroups, which subsequently demonstrated contrasting disease courses. To simplify diagnostic procedure needed for patient stratification, the authors aimed to develop a qPCR test consisting of several of the most significant classifier genes from the whole-blood transcriptome using logistic regression with adaptive elastic net penalty. A list of 39 candidate genes was selected from the top models; however, it was shrunk to 17 genes during the qPCR validation in repeated penalized regression analysis. The 17 genes set (15 informative and 2 reference genes) was further validated in the independent, newly diagnosed group of patients using the qPCR method. The negative predictive value of the established test was very high, which was important for the identification of patients who do not need additional therapy [83]. This is an example of good practices in the development of prediction models that could be easily translated into clinical practice.

Another increasingly explored omics area in IBD is microRNAs [84]. MicroRNAs are more stable than mRNA and easily accessible in blood or urine, which categorize them as promising noninvasive markers for IBD diagnosis. Studies by Hübenthal et al. and Duttagupta et al. used the information on microRNA from peripheral blood to construct prediction models that distinguish between healthy and diseased individuals [67,85]. Hubenthal employed a penalized SVM method for selecting a small set of 16 distinct microRNAs (from a total of 863) which were sufficient for sensitive and specific classification between CD, UC and controls [67]. Duttagupta extracted the signatures of 31 differentially expressed platelet-derived microRNAs that in the SVM model demonstrated high accuracy, specificity and sensitivity in differentiating UC patients from normal individuals [85]. However, the limitations of both studies were small sample sizes and lacked proper independent datasets, which could lead to overfitted models.

Single-cell RNA sequencing technology is increasingly used in IBD research, allowing the detailed analysis of different phenotypes of each cell type. This is very important in IBD research because inflammatory phenotypes of immune cells are enriched in inflamed tissues. In addition, the detection of such cell phenotypes is associated with disease progression and therapy failure, as shown in a study by Martin et al. [69]. In this study, the authors used an unsupervised ML approach—hierarchical clustering—to differentiate between major cell types. The performance of hierarchical clustering using single-cell RNA sequencing data was compared to cytometry analysis, and the results showed a strong correlation between the two methodologies, both in inflamed and uninflamed tissues. Subsequent principal component analysis examined differential cellular subtype frequencies between paired inflamed and uninflamed tissues. The analysis showed that the first two principal components can explain around 73% of variance referred to cellular subtype composition. This approach can help differentiate the cellular composition of inflamed and uninflamed tissue, which could have great potential for clinical application facilitating precise diagnosis, disease localization and therapeutic decisions.

4. Future Perspectives

This review has been focused on machine learning applications in IBD research using genomic and transcriptomic data. Beside this, there is a vast potential among metagenomic, proteomic, epigenomic, metabolomic, and even single-cell transcriptomic data that were and would be explored in machine learning modeling [86–90]. The integration of data from individual IBD-relevant 'omes' is currently considered as the approach that would significantly improve the understanding of IBD pathogenesis and management [1,13,21]. One of the key challenges in this process is to effectively utilize information obtained in omics studies with patients' data stored in electronic medical records (biochemistry tests, various imaging data, symptoms at diagnosis and lifestyle specifics) [38,91]. Machine learning approaches offer the ability to effectively deal with the high dimensionality of these data with the final aim to translate discoveries into clinical practice. Clinical biobanks that gather the multiomics data together with clinical characteristics of patients, such as 1000IBD, RISK and PRISM cohorts, are essential for bringing these up-to-date statistical methodologies to their maximum [92,93]. These synchronized collections of patient metadata provide the raw material for a future significant improvement of precise diagnoses, disease monitoring and personalized treatments. However, to reach these objectives, the prospective validation of AI application should be performed in independent IBD cohorts. The benefits of such methodology are relatively easily demonstrated within one study, but it is much harder to replicate these to independent studies due to high genetic and environmental diversity in human populations [68]. In addition, the variation in clinical decisions and therapeutic protocols, as well as complex nature and heterogeneity of IBD, could affect the successful validation of ML in different cohorts. Since the genetic landscape is population specific, it is particularly important to examine feasibility of omics-based AI models among different populations. Meta-analyses of the existing omic studies can aid identification of reliable and replicable IBD classifiers. Selection of the input data in prediction modelling could be directed by the previous genetic insights and results from meta-analysis. Beside this, the standardization of machine learning techniques as well as practicing transparent, open-sourced and easily reproducible computational research improves the development and replication of the machine learning models in biology. Most of all, randomized clinical trials are needed to determine if these prediction models truly improve clinical outcomes, and if they do, cost effectiveness of their usage compared to standard IBD clinical protocols should be assessed. In addition, the ethical issues that follow an individual's disease predictions should be taken into consideration.

5. Conclusions

IBD is a multifactorial, complex and lifelong disease with varying representation in terms of disease type, age of onset, localization, and severity. Accurate diagnosis and prognosis of the disease followed by the right treatment is of the essence for controlling the disease. Emerging technologies provide the means to collect ever more multiomics data from large cohorts of patients. Instead of relying only on a small number of biomarkers, ML algorithms can employ the big data collected from multiomics analyses coupled with electronic health record data to provide more accurate predictions. Large cohorts enable an opportunity to develop more complex ML models able to capture complex dependencies between features resulting in better predictions and detection of novel biomarkers. Still, before being employed in a clinical setting, predictive models should be rigorously tested in independent cohorts and in the settings of clinical trials to ascertain that this approach can indeed bring benefits to IBD patients in terms of prevention, timely and accurate diagnosis and personalized treatment.

Author Contributions: Conceptualization, B.S., N.K., G.N. and S.P.; writing—original draft preparation, B.S., N.K., G.N., V.G., B.Z. and S.P.; writing—review and editing, B.S., N.K., G.N. and S.P.; supervision, B.S. and S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Education, Science and Technological Development, Republic of Serbia (Grant No. 451-03-9/2021-14/200042).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. De Souza, H.S.P.; Fiocchi, C.; Iliopoulos, D. The IBD interactome: An integrated view of aetiology, pathogenesis and therapy. *Nat. Rev. Gastroenterol. Hepatol.* **2017**, *14*, 739–749. [CrossRef]
2. Ruel, J.; Ruane, D.; Mehandru, S.; Gower-Rousseau, C.; Colombel, J.F. IBD across the age spectrum—Is it the same disease? *Nat. Rev. Gastroenterol. Hepatol.* **2014**, *11*, 88–98. [CrossRef]
3. Ananthakrishnan, A.N.; Shi, H.Y.; Tang, W.; Law, C.C.Y.; Sung, J.J.Y.; Chan, F.K.L.; Ng, S.C. Systematic review and meta-analysis: Phenotype and clinical outcomes of older-onset inflammatory bowel disease. *J. Crohn's Colitis* **2016**, *10*, 1224–1236. [CrossRef] [PubMed]
4. Sartor, R.B. Mechanisms of disease: Pathogenesis of Crohn's disease and ulcerative colitis. *Nat. Clin. Pract. Gastroenterol. Hepatol.* **2006**, *3*, 390–407. [CrossRef] [PubMed]
5. Silverberg, M.S.; Satsangi, J.; Ahmad, T.; Arnott, I.D.R.; Bernstein, C.N.; Brant, S.R.; Caprilli, R.; Colombel, J.F.; Gasche, C.; Geboes, K.; et al. Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: Report of a Working Party of the 2005 Montreal World Congress of Gastroenterology. *Can. J. Gastroenterol.* **2005**, *19* (Suppl. A), 5A–36A. [CrossRef] [PubMed]
6. Danese, S.; Fiocchi, C. Ulcerative Colitis. *N. Engl. J. Med.* **2011**, *365*, 1713–1725. [CrossRef]
7. Xavier, R.J.; Podolsky, D.K. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **2007**, *448*, 427–434. [CrossRef]
8. Levine, A.; Koletzko, S.; Turner, D.; Escher, J.C.; Cucchiara, S.; De Ridder, L.; Kolho, K.L.; Veres, G.; Russell, R.K.; Paerregaard, A.; 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. [CrossRef]
9. Khor, B.; Gardet, A.; Xavier, R.J. Genetics and pathogenesis of inflammatory bowel disease. *Nature* **2011**, *474*, 307–317. [CrossRef]
10. Gevers, D.; Kugathasan, S.; Denson, L.A.; Vázquez-Baeza, Y.; Van Treuren, W.; Ren, B.; Schwager, E.; Knights, D.; Song, S.J.; Yassour, M.; et al. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **2014**, *15*, 382–392. [CrossRef]
11. Cleynen, I.; Boucher, G.; Jostins, L.; Schumm, L.P.; Zeissig, S.; Ahmad, T.; Andersen, V.; Andrews, J.M.; Annese, V.; Brand, S.; et al. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: A genetic association study. *Lancet* **2016**, *387*, 156–167. [CrossRef]
12. Mirkov, M.U.; Verstockt, B.; Cleynen, I. Genetics of inflammatory bowel disease: Beyond NOD2. *Lancet Gastroenterol. Hepatol.* **2017**, *2*, 224–234. [CrossRef]
13. Seyed Tabib, N.S.; Madgwick, M.; Sudhakar, P.; Verstockt, B.; Korcsmaros, T.; Vermeire, S. Big data in IBD: Big progress for clinical practice. *Gut* **2020**, *69*, 1520–1532. [CrossRef]
14. Ogura, Y.; Bonen, D.K.; Inohara, N.; Nicolae, D.L.; Chen, F.F.; Ramos, R.; Britton, H.; Moran, T.; Karaliuskas, R.; Duerr, R.H.; et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **2001**, *411*, 603–606. [CrossRef] [PubMed]
15. McGovern, D.P.B.; Kugathasan, S.; Cho, J.H. Genetics of Inflammatory Bowel Diseases. *Gastroenterology* **2015**, *149*, 1163–1176.e2. [CrossRef]
16. Glocker, E.-O.; Kotlarz, D.; Boztug, K.; Gertz, E.M.; Schäffer, A.A.; Noyan, F.; Perro, M.; Diestelhorst, J.; Allroth, A.; Murugan, D.; et al. Inflammatory Bowel Disease and Mutations Affecting the Interleukin-10 Receptor. *N. Engl. J. Med.* **2009**, *361*, 2033–2045. [CrossRef] [PubMed]
17. Liu, J.Z.; Van Sommeren, S.; Huang, H.; Ng, S.C.; Alberts, R.; Takahashi, A.; Ripke, S.; Lee, J.C.; Jostins, L.; Shah, T.; et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **2015**, *47*, 979–986. [CrossRef]
18. Stappenbeck, T.S.; Rioux, J.D.; Mizoguchi, A.; Saitoh, T.; Huett, A.; Darfeuille-Michaud, A.; Wileman, T.; Mizushima, N.; Carding, S.; Akira, S.; et al. Crohn disease: A current perspective on genetics, autophagy and immunity. *Autophagy* **2011**, *7*, 355–374. [CrossRef]
19. Na, Y.R.; Stakenborg, M.; Seok, S.H.; Matteoli, G. Macrophages in intestinal inflammation and resolution: A potential therapeutic target in IBD. *Nat. Rev. Gastroenterol. Hepatol.* **2019**, *16*, 531–543. [CrossRef]
20. Salas, A.; Hernandez-Rocha, C.; Duijvestein, M.; Faubion, W.; McGovern, D.; Vermeire, S.; Vetrano, S.; Vande Casteele, N. JAK–STAT pathway targeting for the treatment of inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* **2020**, *17*, 323–337. [CrossRef]
21. Furey, T.S.; Sethupathy, P.; Sheikh, S.Z. Redefining the IBDs using genome-scale molecular phenotyping. *Nat. Rev. Gastroenterol. Hepatol.* **2019**, *16*, 296–311. [CrossRef]
22. Mishra, R.; Dhawan, P.; Srivastava, A.S.; Singh, A.B. Inflammatory bowel disease: Therapeutic limitations and prospective of the stem cell therapy. *World J. Stem Cells* **2020**, *12*, 1050–1066. [CrossRef]

23. Henderson, P.; Van Limbergen, J.E.; Schwarze, J.; Wilson, D.C. Function of the intestinal epithelium and its dysregulation in inflammatory bowel disease. *Inflamm. Bowel Dis.* **2011**, *17*, 382–395. [CrossRef]
24. Dyson, J.K.; Rutter, M.D. Colorectal cancer in inflammatory bowel disease: What is the real magnitude of the risk? *World J. Gastroenterol.* **2012**, *18*, 3839–3848. [CrossRef]
25. Tran, V.; Shammas, R.M.; Sauk, J.S.; Padua, D. Evaluating tofacitinib citrate in the treatment of moderate-to-severe active ulcerative colitis: Design, development and positioning of therapy. *Clin. Exp. Gastroenterol.* **2019**, *12*, 179–191. [CrossRef] [PubMed]
26. Rogler, G. Gastrointestinal and liver adverse effects of drugs used for treating IBD. *Best Pract. Res. Clin. Gastroenterol.* **2010**, *24*, 157–165. [CrossRef] [PubMed]
27. Adegbola, S.O.; Sahnan, K.; Warusavitarne, J.; Hart, A.; Tozer, P. Anti-TNF therapy in Crohn's disease. *Int. J. Mol. Sci.* **2018**, *19*, 2244. [CrossRef] [PubMed]
28. Kohli, A.; Holzwanger, E.A.; Levy, A.N. Emerging use of artificial intelligence in inflammatory bowel disease. *World J. Gastroenterol.* **2020**, *26*, 6923–6928. [CrossRef] [PubMed]
29. Brookes, M.J.; Green, J.R.B. Maintenance of remission in Crohn's disease: Current and emerging therapeutic options. *Drugs* **2004**, *64*, 1069–1089. [CrossRef] [PubMed]
30. Jostins, L.; Ripke, S.; Weersma, R.K.; Duerr, R.H.; McGovern, D.P.; Hui, K.Y.; Lee, J.C.; Philip Schumm, L.; Sharma, Y.; Anderson, C.A.; et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **2012**, *491*, 119–124. [CrossRef]
31. Wu, F.; Zhang, S.; Dassopoulos, T.; Harris, M.L.; Bayless, T.M.; Meltzer, S.J.; Brant, S.R.; Kwon, J.H. Identification of microRNAs associated with ileal and colonic Crohn's disease. *Inflamm. Bowel Dis.* **2010**, *16*, 1729–1738. [CrossRef]
32. Dulai, P.S.; Singh, S.; Vande Casteele, N.; Boland, B.S.; Rivera-Nieves, J.; Ernst, P.B.; Eckmann, L.; Barrett, K.E.; Chang, J.T.; Sandborn, W.J. Should We Divide Crohn's Disease Into Ileum-Dominant and Isolated Colonic Diseases? *Clin. Gastroenterol. Hepatol.* **2019**, *17*, 2634–2643. [CrossRef] [PubMed]
33. Häsler, R.; Feng, Z.; Bäckdahl, L.; Spehlmann, M.E.; Franke, A.; Teschendorff, A.; Rakyan, V.K.; Down, T.A.; Wilson, G.A.; Feber, A.; et al. A functional methylome map of ulcerative colitis. *Genome Res.* **2012**, *22*, 2130–2137. [CrossRef] [PubMed]
34. Haberman, Y.; Tickle, T.L.; Dexheimer, P.J.; Kim, M.O.; Tang, D.; Karns, R.; Baldassano, R.N.; Noe, J.D.; Rosh, J.; Markowitz, J.; et al. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J. Clin. Investig.* **2014**, *124*, 3617–3633. [CrossRef]
35. Peck, B.C.E.; Weiser, M.; Lee, S.E.; Gipson, G.R.; Iyer, V.B.; Sartor, R.B.; Herfarth, H.H.; Long, M.D.; Hansen, J.J.; Isaacs, K.L.; et al. MicroRNAs classify different disease behavior phenotypes of Crohn's disease and may have prognostic utility. *Inflamm. Bowel Dis.* **2015**, *21*, 2178–2187. [CrossRef] [PubMed]
36. Marigorta, U.M.; Denson, L.A.; Hyams, J.S.; Mondal, K.; Prince, J.; Walters, T.D.; Griffiths, A.; Noe, J.D.; Crandall, W.V.; Rosh, J.R.; et al. Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nat. Genet.* **2017**, *49*, 1517–1521. [CrossRef]
37. Romagnoni, A.; Jégou, S.; Van Steen, K.; Wainrib, G.; Hugot, J.P.; Peyrin-Biroulet, L.; Chamaillard, M.; Colombel, J.F.; Cottone, M.; D'Amato, M.; et al. Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. *Sci. Rep.* **2019**, *9*, 10351. [CrossRef]
38. Gubatan, J.; Levitte, S.; Patel, A.; Balabanis, T.; Wei, M.T.; Sinha, S.R. Artificial intelligence applications in inflammatory bowel disease: Emerging technologies and future directions. *World J. Gastroenterol.* **2021**, *27*, 1920–1935. [CrossRef]
39. Stankovic, B.; Dragasevic, S.; Popovic, D.; Zukic, B.; Kotur, N.; Sokic-Milutinovic, A.; Alempijevic, T.; Lukic, S.; Milosavljevic, T.; Nikcevic, G.; et al. Variations in inflammatory genes as molecular markers for prediction of inflammatory bowel disease occurrence. *J. Dig. Dis.* **2015**, *16*, 723–733. [CrossRef]
40. Waljee, A.K.; Lipson, R.; Wiitala, W.L.; Zhang, Y.; Liu, B.; Zhu, J.; Wallace, B.; Govani, S.M.; Stidham, R.W.; Hayward, R.; et al. Predicting Hospitalization and Outpatient Corticosteroid Use in Inflammatory Bowel Disease Patients Using Machine Learning. *Inflamm. Bowel Dis.* **2017**, *24*, 45–53. [CrossRef]
41. Waljee, A.K.; Joyce, J.C.; Wang, S.; Saxena, A.; Hart, M.; Zhu, J.; Higgins, P.D.R. Algorithms outperform metabolite tests in predicting response of patients with inflammatory bowel disease to thiopurines. *Clin. Gastroenterol. Hepatol.* **2010**, *8*, 143–150. [CrossRef]
42. Waljee, A.K.; Liu, B.; Sauder, K.; Zhu, J.; Govani, S.M.; Stidham, R.W.; Higgins, P.D.R. Predicting corticosteroid-free endoscopic remission with vedolizumab in ulcerative colitis. *Aliment. Pharmacol. Ther.* **2018**, *47*, 763–772. [CrossRef]
43. Waljee, A.K.; Liu, B.; Sauder, K.; Zhu, J.; Govani, S.M.; Stidham, R.W.; Higgins, P.D.R. Predicting Corticosteroid-Free Biologic Remission with Vedolizumab in Crohn's Disease. *Inflamm. Bowel Dis.* **2018**, *24*, 1185–1192. [CrossRef]
44. Mossotto, E.; Ashton, J.J.; Coelho, T.; Beattie, R.M.; MacArthur, B.D.; Ennis, S. Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. *Sci. Rep.* **2017**, *7*, 2427. [CrossRef] [PubMed]
45. Schneider, N.; Sohrabi, K.; Schneider, H.; Zimmer, K.-P.; Fischer, P.; de Laffolie, J. Machine Learning Classification of Inflammatory Bowel Disease in Children Based on a Large Real-World Pediatric Cohort CEDATA-GPGE®Registry. *Front. Med.* **2021**, *8*, 666190. [CrossRef] [PubMed]
46. Ozawa, T.; Ishihara, S.; Fujishiro, M.; Saito, H.; Kumagai, Y.; Shichijo, S.; Aoyama, K.; Tada, T. Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. *Gastrointest. Endosc.* **2019**, *89*, 416–421.e1. [CrossRef]

47. Stidham, R.W.; Liu, W.; Bishu, S.; Rice, M.D.; Higgins, P.D.R.; Zhu, J.; Nallamothu, B.K.; Waljee, A.K. Performance of a Deep Learning Model vs Human Reviewers in Grading Endoscopic Disease Severity of Patients With Ulcerative Colitis. *JAMA Netw. Open* **2019**, *2*, e193963. [CrossRef] [PubMed]
48. Brownlee, J. Master Machine Learning Algorithms: Discover how they work and implement them from scratch. In *Machine Learning Mastery*, 5th ed.; Cambridge University Press: Cambridge, UK, 2014.
49. Isakov, O.; Dotan, I.; Ben-Shachar, S. Machine Learning-Based Gene Prioritization Identifies Novel Candidate Risk Genes for Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* **2017**, *23*, 1516–1523. [CrossRef]
50. Li, H.; Lai, L.; Shen, J. Development of a susceptibility gene based novel predictive model for the diagnosis of ulcerative colitis using random forest and artificial neural network. *Aging (Albany. NY)* **2020**, *12*, 20471–20482. [CrossRef]
51. Pal, L.R.; 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.* **2017**, *38*, 1225–1234. [CrossRef]
52. Cushing, K.C.; McLean, R.; McDonald, K.G.; Gustafsson, J.K.; Knoop, K.A.; Kulkarni, D.H.; Sartor, R.B.; Newberry, R.D. Predicting risk of postoperative disease recurrence in Crohn's disease: Patients with indolent Crohn's disease have distinct whole transcriptome profiles at the time of first surgery. *Inflamm. Bowel Dis.* **2019**, *25*, 180–193. [CrossRef]
53. Khorasani, H.M.; Usefi, H.; Peña-Castillo, L. Detecting ulcerative colitis from colon samples using efficient feature selection and machine learning. *Sci. Rep.* **2020**, *10*, 13744. [CrossRef]
54. Shai, S.-S.; Shai, B.-D. *UNDERSTANDING MACHINE LEARNING—From Theory to Algorithms*; Cambridge University Press: New York, NY, USA, 2014; pp. 114–123.
55. Melo, F. Area under the ROC Curve. In *Encyclopedia of Systems Biology*; Springer New York: New York, NY, USA, 2013; pp. 38–39.
56. Smith, A.M.; Walsh, J.R.; Long, J.; Davis, C.B.; Henstock, P.; Hodge, M.R.; Maciejewski, M.; Mu, X.J.; Ra, S.; Zhao, S.; et al. Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinform.* **2020**, *21*, 119. [CrossRef]
57. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer New York: New York, NY, USA, 2009; ISBN 978-0-387-84857-0.
58. Kobayashi, S.; Saltz, J.H.; Yang, V.W. State of machine and deep learning in histopathological applications in digestive diseases. *World J. Gastroenterol.* **2021**, *27*, 2545–2575. [CrossRef] [PubMed]
59. Heller, K.A.; Ghahramani, Z. Bayesian hierarchical clustering. In Proceedings of the 22nd International Conference on Machine Learning—ICML, Bonn, Germany, 7–11 August 2005; ACM Press: New York, NY, USA, 2005; pp. 297–304.
60. Abraham, G.; Kowalczyk, A.; Zobel, J.; Inouye, M. Performance and Robustness of Penalized and Unpenalized Methods for Genetic Prediction of Complex Human Disease. *Genet. Epidemiol.* **2013**, *37*, 184–195. [CrossRef]
61. Han, L.; Maciejewski, M.; Brockel, C.; Gordon, W.; Snapper, S.B.; Korzenik, J.R.; Afzelius, L.; Altman, R.B. A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. *Bioinformatics* **2018**, *34*, 985–993. [CrossRef] [PubMed]
62. Yuan, F.; Zhang, Y.-H.; Kong, X.-Y.; Cai, Y.-D. Identification of Candidate Genes Related to Inflammatory Bowel Disease Using Minimum Redundancy Maximum Relevance, Incremental Feature Selection, and the Shortest-Path Approach. *BioMed Res. Int.* **2017**, *2017*, 5741948. [CrossRef] [PubMed]
63. Raimondi, D.; Simm, J.; Arany, A.; Fariselli, P.; Cleynen, I.; Moreau, Y. 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. [CrossRef]
64. Wang, Y.; Miller, M.; Astrakhan, Y.; Petersen, B.S.; Schreiber, S.; Franke, A.; Bromberg, Y. Identifying Crohn's disease signal from variome analysis. *Genome Med.* **2019**, *11*, 59. [CrossRef]
65. Chen, G.B.; Lee, S.H.; Montgomery, G.W.; Wray, N.R.; Visscher, P.M.; Gearry, R.B.; Lawrance, I.C.; Andrews, J.M.; Bampton, P.; Mahy, G.; et al. Performance of risk prediction for inflammatory bowel disease based on genotyping platform and genomic risk score method. *BMC Med. Genet.* **2017**, *18*, 94. [CrossRef]
66. Wei, Z.; Wang, W.; Bradfield, J.; Li, J.; Cardinale, C.; Frackelton, E.; Kim, C.; Mentch, F.; Van Steen, K.; Visscher, P.M.; et al. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am. J. Hum. Genet.* **2013**, *92*, 1008–1012. [CrossRef]
67. Hübenthal, M.; Hemmrich-Stanisak, G.; Degenhardt, F.; Szymczak, S.; Du, Z.; Elsharawy, A.; Keller, A.; Schreiber, S.; Franke, A. Sparse modeling reveals miRNA signatures for diagnostics of inflammatory bowel disease. *PLoS ONE* **2015**, *10*, e0140155. [CrossRef]
68. Zarringhalam, K.; Enayetallah, A.; Reddy, P.; Ziemek, D. Robust clinical outcome prediction based on Bayesian analysis of transcriptional profiles and prior causal networks. *Bioinformatics* **2014**, *30*, i69–i77. [CrossRef]
69. Martin, J.C.; Chang, C.; Boschetti, G.; Ungaro, R.; Giri, M.; Grout, J.A.; Gettler, K.; Chuang, L.-S.; Nayar, S.; Greenstein, A.J.; et al. Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* **2019**, *178*, 1493–1508.e20. [CrossRef]
70. Daneshjou, R.; Wang, Y.; Bromberg, Y.; Bovo, S.; Martelli, P.L.; Babbi, G.; Di Lena, P.; Casadio, R.; Edwards, M.; Gifford, D.; et al. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Hum. Mutat.* **2017**, *38*, 1182–1192. [CrossRef] [PubMed]

71. Giollo, M.; Jones, D.T.; Carraro, M.; Leonardi, E.; Ferrari, C.; Tosatto, S.C.E. Crohn disease risk prediction—Best practices and pitfalls with exome data. *Hum. Mutat.* **2017**, *38*, 1193–1200. [CrossRef] [PubMed]
72. Pal, L.R.; Yu, C.H.; Mount, S.M.; Moult, J. Insights from GWAS: Emerging landscape of mechanisms underlying complex trait disease. *BMC Genom.* **2015**, *16*, 54. [CrossRef] [PubMed]
73. Yu, W.; Clyne, M.; Khoury, M.J.; Gwinn, M. Phenopedia and genopedia: Disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics* **2009**, *26*, 145–146. [CrossRef]
74. Jostins, L.; Barrett, J.C. Genetic risk prediction in complex disease. *Hum. Mol. Genet.* **2011**, *20*, 182–188. [CrossRef]
75. Wray, N.R.; Yang, J.; Goddard, M.E.; Visscher, P.M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet.* **2010**, *6*, e1000864. [CrossRef]
76. Liu, J.Z.; Anderson, C.A. Genetic studies of Crohn's disease: Past, present and future. *Best Pract. Res. Clin. Gastroenterol.* **2014**, *28*, 373–386. [CrossRef]
77. Cleynen, I.; González, J.R.; Figueroa, C.; Franke, A.; McGovern, D.; Bortlík, M.; Crusius, B.J.A.; Vecchi, M.; Artieda, M.; Szczypiorska, M.; 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*, 1556–1565. [CrossRef] [PubMed]
78. Cleynen, I.; Vermeire, S. The genetic architecture of inflammatory bowel disease: Past, present and future. *Curr. Opin. Gastroenterol.* **2015**, *31*, 456–463. [CrossRef] [PubMed]
79. Collado-Torres, L.; Nellore, A.; Kammers, K.; Ellis, S.E.; Taub, M.A.; Hansen, K.D.; Jaffe, A.E.; Langmead, B.; Leek, J.T. Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **2017**, *35*, 319–321. [CrossRef] [PubMed]
80. Stankovic, B.; Dragasevic, S.; Klaassen, K.; Kotur, N.; Srzentic Drazilov, S.; Zukic, B.; Sokic Milutinovic, A.; Milovanovic, T.; Lukic, S.; Popovic, D.; et al. Exploring inflammatory and apoptotic signatures in distinct Crohn's disease phenotypes: Way towards molecular stratification of patients and targeted therapy. *Pathol. Res. Pract.* **2020**, *216*, 152945. [CrossRef]
81. Anderson, C.A.; Boucher, G.; Lees, C.W.; Franke, A.; D'Amato, M.; Taylor, K.D.; Lee, J.C.; Goyette, P.; Imielinski, M.; Latiano, A.; et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **2011**, *43*, 246–252. [CrossRef] [PubMed]
82. Conrad, K.; Roggenbuck, D.; Laass, M.W. Diagnosis and classification of ulcerative colitis. *Autoimmun. Rev.* **2014**, *13*, 463–466. [CrossRef] [PubMed]
83. Biasci, D.; Lee, J.C.; Noor, N.M.; Pombal, D.R.; Hou, M.; Lewis, N.; Ahmad, T.; Hart, A.; Parkes, M.; McKinney, E.F.; et al. A blood-based prognostic biomarker in IBD. *Gut* **2019**, *68*, 1386–1395. [CrossRef] [PubMed]
84. Cao, B.; Zhou, X.; Ma, J.; Zhou, W.; Yang, W.; Fan, D.; Hong, L. Role of MiRNAs in Inflammatory Bowel Disease. *Dig. Dis. Sci.* **2017**, *62*, 1426–1438. [CrossRef]
85. Duttagupta, R.; DiRienzo, S.; Jiang, R.; Bowers, J.; Gollub, J.; Kao, J.; Kearney, K.; Rudolph, D.; Dawany, N.B.; Showe, M.K.; et al. Genome-wide maps of circulating miRNA biomarkers for Ulcerative Colitis. *PLoS ONE* **2012**, *7*, e31241. [CrossRef]
86. Douglas, G.M.; Hansen, R.; Jones, C.M.A.; Dunn, K.A.; Comeau, A.M.; Bielawski, J.P.; Tayler, R.; El-Omar, E.M.; Russell, R.K.; Hold, G.L.; et al. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. *Microbiome* **2018**, *6*, 13. [CrossRef]
87. Reiman, D.; Layden, B.T.; Dai, Y. MiMeNet: Exploring microbiome-metabolome relationships using neural networks. *PLoS Comput. Biol.* **2021**, *17*, e1009021. [CrossRef] [PubMed]
88. Weiser, M.; Simon, J.M.; Kochar, B.; Tovar, A.; Israel, J.W.; Robinson, A.; Gipson, G.R.; Schaner, M.S.; Herfarth, H.H.; Sartor, R.B.; et al. Molecular classification of Crohn's disease reveals two clinically relevant subtypes. *Gut* **2018**, *67*, 36–42. [CrossRef]
89. Ungaro, R.C.; Hu, L.; Ji, J.; Nayar, S.; Kugathasan, S.; Denson, L.A.; Hyams, J.; Dubinsky, M.C.; Sands, B.E.; Cho, J.H. Machine learning identifies novel blood protein predictors of penetrating and stricturing complications in newly diagnosed paediatric Crohn's disease. *Aliment. Pharmacol. Ther.* **2021**, *53*, 281–290. [CrossRef] [PubMed]
90. Smillie, C.S.; Biton, M.; Ordovas-Montanes, J.; Sullivan, K.M.; Burgin, G.; Graham, D.B.; Herbst, R.H.; Rogel, N.; Slyper, M.; Waldman, J.; et al. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **2019**, *178*, 714–730.e22. [CrossRef] [PubMed]
91. Stafford, I.S.; Kellermann, M.; Mossotto, E.; Beattie, R.M.; MacArthur, B.D.; Ennis, S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ Digit. Med.* **2020**, *3*, 1–11. [CrossRef] [PubMed]
92. Imhann, F.; Van Der Velde, K.J.; Barbieri, R.; Alberts, R.; Voskuil, M.D.; Vich Vila, A.; Collij, V.; Spekhorst, L.M.; Der Sloot Kwj, V.; Peters, V.; et al. The 1000IBD project: Multi-omics data of 1000 inflammatory bowel disease patients; Data release 1. *BMC Gastroenterol.* **2019**, *19*, 5. [CrossRef]
93. Proctor, L.M.; DiGiacomo, N.D.; Fettweis, J.M.; Jefferson, K.K.; Strauss, J.F.; Rubens, C.E.; Brooks, J.P.; Girerd, P.P.; Huang, B.; Serrano, M.G. The Integrative Human Microbiome Project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* **2014**, *16*, 276–289. [CrossRef]