

Article

Gene Identification in Inflammatory Bowel Disease via a Machine Learning Approach

Gerardo Alfonso Perez *  and Raquel Castillo 

Biocomp Group, Institute of Advanced Materials (INAM), Universitat Jaume I, 12071 Castello, Spain

* Correspondence: ga284@cantab.net

Abstract: Inflammatory bowel disease (IBD) is an illness with increasing prevalence, particularly in emerging countries, which can have a substantial impact on the quality of life of the patient. The illness is rather heterogeneous with different evolution among patients. A machine learning approach is followed in this paper to identify potential genes that are related to IBD. This is done by following a Monte Carlo simulation approach. In total, 23 different machine learning techniques were tested (in addition to a base level obtained using artificial neural networks). The best model identified 74 genes selected by the algorithm as being potentially involved in IBD. IBD seems to be a polygenic illness, in which environmental factors might play an important role. Following a machine learning approach, it was possible to obtain a classification accuracy of 84.2% differentiating between patients with IBD and control cases in a large cohort of 2490 total cases. The sensitivity and specificity of the model were 82.6% and 84.4%, respectively. It was also possible to distinguish between the two main types of IBD: (1) Crohn's disease and (2) ulcerative colitis.

Keywords: inflammatory bowel disease; Crohn's disease; ulcerative colitis



Citation: Alfonso Perez, G.; Castillo, R. Gene Identification in Inflammatory Bowel Disease via a Machine Learning Approach. *Medicina* **2023**, *59*, 1218. <https://doi.org/10.3390/medicina59071218>

Academic Editors: Wisit Cheungpasitporn, Charat Thongprayoon, Pattharawin Pattharanitima and Wisit Kaewput

Received: 1 June 2023
Revised: 24 June 2023
Accepted: 26 June 2023
Published: 28 June 2023



Copyright: © 2023 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

In this paper, the genetic expression signature of inflammatory bowel disease is analyzed using machine learning techniques. Inflammatory bowel disease (IBD) is a chronic [1] inflammatory disease, whose cause remains unclear. Patients can show an array of different symptoms. According to the Mayo Clinic, some of the most common symptoms associated with inflammatory bowel disease include pain, diarrhea, fatigue, cramps, blood present in stools and weight loss. Extraintestinal symptoms appear in approximately 24% of patients [2]. Patients can also have very different evolution and responses to treatments.

Another interesting characteristic of this illness, so far without a good explanation, is that it tends to have a higher incidence and prevalence in urban areas [3] compared to rural areas, perhaps suggesting a link to lifestyles. The incidence of IBD has been increasing [4]. Inflammatory bowel disease is becoming an increasingly important health problem [5]. Developing and newly industrialized countries are seeing a particularly rapid increase in the incidence of the illness [6]. The reasons behind this increase remain unclear. It might be related to changes in dietary habits or exposure to pollutants, but there are currently, to the best of our knowledge, no definitive data to prove it. It is also likely that the illness is being detected earlier in those countries as their healthcare infrastructure develops. Nevertheless, environmental factors appear to play a role in the illness. IBD increases the chances of developing other illnesses, such as colorectal cancer [7] and osteoporosis [8]. More than 7% of patients with IBD develop osteoporosis [8]. Additionally, IBD can have a very significant impact on the quality of life of the patient and can make normal activities, such as working, challenging in some severe cases.

One of the main theories of the cause of IBD is that it is an abnormal immune response in genetically predisposed individuals, triggered by some external factor such as a virus or bacteria [9,10]. Cytokines appear to play an important role in IBD [11]. Lifestyle

factors, such as stress, smoking and diet [12], have also been identified in the literature as having a role in the illness [13]. The illness results in a defective regulation of the mucosa [14]. Tamboli et al. [15] specifically mentioned intestinal bacteria as a major factor in the initial stages of the disease. Chang [16] concluded that the two causative agents are (1) abnormal immune response in the gastrointestinal mucosa and (2) alterations in the gut microbiome [17]. The two major forms of IBD are ulcerative colitis (UC) and Crohn's disease (CD) [18]. A visual representation of UC and CD is shown in Figure 1.

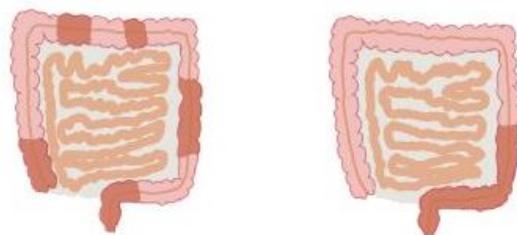


Figure 1. Visual representation of Crohn's disease (left) and ulcerative colitis (right). It can be seen some of the usual areas involved in UC and CD. It should be noted that there is substantial variation among patients.

IBD appears to have a genetic component. Loddo and Romano [19] mentioned that approximately 15% of the patients with Crohn's disease have a family member with the same condition. They also mentioned a 50% concordance in monozygotic twins. Bernard and Ramnik [20] concluded that genes help regulate the complex interaction between microbial and environmental factors. Another indication of a genetic component in the disease is that some ethnic groups, such as Ashkenazim, have higher incidence and prevalence [21]. Some authors, such as McGovern et al. [22], highlighted the issue that a large amount of the existing literature focuses on individuals of European ancestry. This is especially important in an illness such as IBD, in which ethnicity seems to play an important role not only in terms of prevalence but also in terms of early onset, reaction to the treatment and severity of the illness. A schematic representation of the interaction between genetic predisposition and environmental factors is shown in Figure 2. The underlying mechanics of this interaction between genetic predisposition and environmental factors remain not well understood.

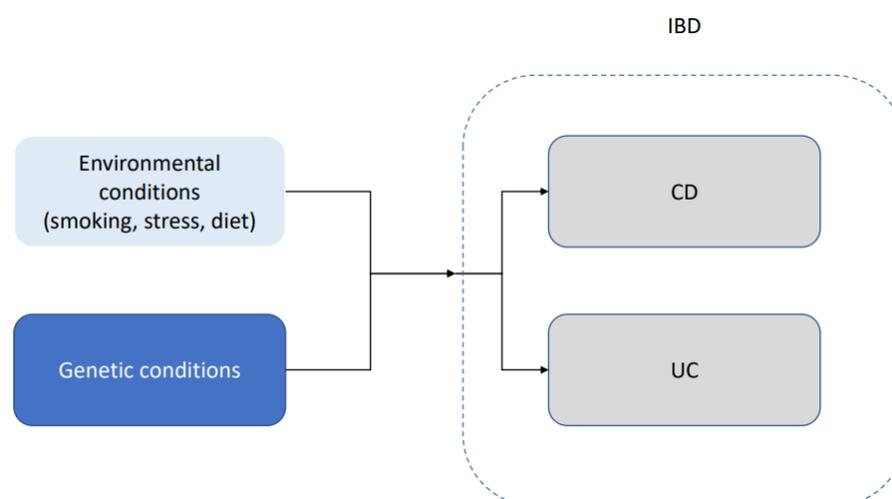


Figure 2. Schematic representation of the interaction between genetic predisposition and environmental factors in ulcerative colitis (UC) and Crohn's disease (CD). IBD, in both of its main forms, is likely caused by a combination of underlying genetic conditions and environmental conditions.

There have been many developments in the genetics of IBD, but despite the identification of some genes, the underlying process remains not well understood. The evidence points to a process in which multiple genes are involved (polygenic) [23,24]. Cho and Abraham [25] cited the well-known Nod2 (CARD15) polymorphism association with Crohn's disease. This gene is located in chromosome 16 and has been mentioned by multiple authors [26]. Katuka et al. [27] mentioned that in Japan, the NUDT15 polymorphism is routinely tested before administering thiopurine to inflammatory bowel disease patients. Mathew and Lewis [28] studied genes in chromosome 5q31n 6p21 and 19p. Achkar and Duerr [29] identified IL23R and ATG16L1 as being involved in CD. These two genes are frequently mentioned in the existing literature [30]. Stoll et al. [31] identified DLG5, while Cleynen et al. [32] identified 163 susceptibility loci for IBD. Ahmad et al. mentioned that CD and UC are related diseases that share some but not all the susceptibility genes [33]. Inflammatory bowel disease is a chronic disease that typically requires lifelong medication [34]. Given the heterogeneity in the illness, it is not surprising that there are multiple treatment options with different levels of expected success.

Machine learning techniques are increasingly popular in medicine with applications in many different types of illness [35–37]. There has been some interesting research applying machine learning techniques in the context of inflammatory bowel disease [38–40]. This has been in part due to the large amount of data generated experimentally [41] and the need to come up with appropriate techniques to analyze such a large quantity of data. For instance, Wei et al. [42] used GWAS data to carry out a risk assessment of patients with ulcerative colitis or Crohn's disease. Isakov et al. [43] identified 67 genes using machine learning techniques related to IBD. Coelho et al. [44] also used machine learning techniques, but their analysis covers pediatric patients, who have some characteristics different from the usual adult case. The same group of authors published another interesting paper [38] using three different machine learning techniques and endoscopic data, achieving an accuracy of 71.0%, 76.9% and 82.7% respectively. The work of Smolander et al. [45] is another interesting paper analyzing gene expression, using machine learning techniques in the context of complex disorders. Some authors, such as Stankvic et al. [46], mentioned that despite an increase in the use of machine learning techniques in IBD, the understanding of the illness remains incomplete.

One of the main objectives of this article is trying to identify genes that are relevant in the context of inflammatory bowel disease using machine learning techniques. The genes are chosen by selecting those genes with a gene expression level that is empirically useful to distinguish between control individuals and patients with IBD. The details of this process will be explained in the next section, but it is based on using different machine learning techniques (classification purposes) in combination with Monte Carlo simulations for the selection of genes. Another objective of this article is to be able to identify appropriate genes differentiating between Crohn's disease and ulcerative colitis using a similar approach than when distinguishing between healthy and IBD patients.

2. Materials and Methods

The dataset was retrieved from the Gene Expression Omnibus. The identification number is GSE 193677 [47]. The data include 2490 total cases. Of these 2490 cases, 461 are controls cases, while 2029 are individuals with adult inflammatory bowel disease (IBD). Of those 2029, a slight majority of 1157 have Crohn's disease while 872 have ulcerative colitis. The average age of the patient is 44.9 years, with a range from 19 to 82 years old. A histogram showing the age distribution is shown in Figure 3. There are 1174 female and 1316 male cases. Tissue biopsies were obtained in the right colon, left colon, transverse, rectum, ileum, sigmoid and cecum. The number of cases for each of this regions is summarized below in Table 1. The data consist of gene expression profiling by high throughput sequencing obtained using the Illumina HiSeq 2500 (Illumina, Inc. San Diego, CA, USA). There are 56,632 expression profiling data per patient.

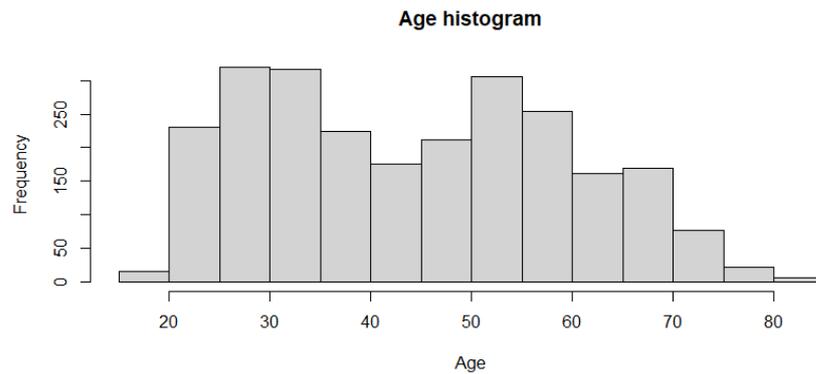


Figure 3. Histogram describing the age of the patients. The range is from 19 to 82 years old.

Table 1. Biopsies (tissue areas).

Area	Cases
Rectum	904
Left colon	180
Right colon	252
Ileum	672
Transverse	90
Sigmoid	163
Cecum	229

The data were divided into two subgroups, a training dataset and a testing dataset. Ψ_{Tr} denotes the training dataset and Ψ_{Ts} the testing dataset. The training and testing datasets contain approximately 80% and 20% of all the cases, respectively. Each column represents a patient. The division into a training and a testing dataset was carried out in a randomized way to try to avoid introducing biases in the analysis. The first row in each dataset contains a numerical classifier identifying the subject as a control or patient (UD or CD) as shown in Equation (1):

$$\forall j \in [1, n], \Phi_j = \{Control = 0, UC = CD = 1\} \tag{1}$$

with n being the total number of cases. An example, for clarity purposes, can be seen in Equation (2):

$$\Phi = \{\Phi_1, \Phi_2, \dots, \Phi_n\} = \{1, 0, \dots, 1\} \tag{2}$$

The following two rows contain the age (a), see Equations (3) and (4), and the gender (S), see Equations (5) and (6), of each individual, respectively:

$$\forall j \in [1, n], a_j = \{x_j\} \ x \in \mathbb{R} \tag{3}$$

$$a = \{age\} = \{a_1, a_2, \dots, a_n\} = \{47, 52, \dots, 61\} \tag{4}$$

$$\forall j \in [1, n], S_j = \{Female = 0, Male = 1\} \tag{5}$$

$$S = \{gender\} = \{S_1, S_2, \dots, S_m\} = \{0, 1, \dots, 1\} \tag{6}$$

In a similar way, the following row contains the region for the biopsy. All the other rows contain gene expression data (see Equations (7) and (8)):

$$\forall j \in [1, n], \forall k \in [1, m], g_{kj} = \{Z_{kj}\} \ Z \in \mathbb{R} \tag{7}$$

$$G_k = \{g_{kj}\} = \{18, 241, \dots, 132\} \tag{8}$$

where k is the index for each row. An example, for visualization purposes, of the data can be seen in Equation (9):

$$\Psi_{Tr} = \begin{pmatrix} 0 & 1 & 2 & 0 & \dots \\ 60 & 45 & 35 & 55 & \dots \\ 0 & 0 & 1 & 1 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ 80 & 30 & 55 & 40 & \dots \\ \vdots & \vdots & & \vdots & \end{pmatrix} \tag{9}$$

As a first step, the correlation $C_0(c, d)$ between the categorical data representing the classification group (control or IBD) and each row is calculated (Equations (10)):

$$\forall k \in [1, m], C_0 = C_0(\Phi, G_k) \tag{10}$$

Therefore, C_0 is a vector with m components. From this mapping, the highest $q\%$ ($0 \leq q \leq 100$) is selected among these m values. Hence, there is a reduction in the dimension of the vector (Equation (11)):

$$C_0(dim = m) \Rightarrow C_0^*(dim = m < k) \tag{11}$$

This step is performed in an attempt to include the factors that are potentially able to generate an accurate model while filtering out potential noise (not all genes are involved in inflammatory bowel disease). In other words, it is an attempt to filter out noise from genes than have no biological impact on the disease but that can lead the model to find spurious relationships given the large amount of data. The above-mentioned step is carried out only with the training dataset (containing approximately 80% of the cases). After this step, when the genes have already been selected, then all the other genes will be excluded from both the training and the testing dataset. In this way, it is possible to carry out a filtering of the initial gene list. A selection of 23 machine learning techniques was selected; see Table 2. Ten times cross validation was carried out (training dataset).

Table 2. Selected machine learning algorithms.

Algorithm	Algorithm
Complex Tree	Fine KNN
Medium Tree	Medium KNN
Simple Tree	Coarse KNN
Linear Discriminant	Cosine KNN
Quadratic Discriminant	Cubic KNN
Logistic Regression	Weighted KNN
Linear SVM	Boosted Trees
Quadratic SVM	Bagged Trees
Cubic SVM	Subspace Discriminant
Fine Gaussian SVM	Subspace KNN
Medium Gaussian SVM	RUSBoosted Trees
Coarse Gaussian SVM	

The artificial neural network (ANN) is a well-known machine learning algorithm. Given its versatility and wide use, this technique is used to determine a baseline classification accuracy, against which the other techniques are compared. In the ANN approach, it is necessary to carry out hyperparameter optimization. One of the key parameters to optimize is the number of layers in the ANN. This is achieved by carrying out simulations

from 1 to 1000 layers and the related accuracy estimated. Unless explicitly mentioned, the accuracy (and other measures of the goodness of the fit) is that of the testing dataset (not used during the training phase). In this way, for each configuration γ ($\gamma = \{1, \dots, 1000\}$), an accuracy A_{nn} measure is estimated (A_{nn}^γ). Then, the best model (\bar{A}_{nn}) is selected as

$$\bar{A}_{nn}(\gamma) = \text{sup}(A_c^\gamma) \tag{12}$$

This is the baseline model. For each machine learning techniques, the model is trained with the training dataset, and then an accuracy estimate is obtained, and the best model $\bar{A}(\lambda)$ is selected (Equation (13)). The training and model selection (gene selection) is entirely performed with the training dataset. After the model is selected (including the genes), the accuracy and other metrics are expressed in terms of the testing dataset (not used for training or model selection):

$$\bar{A}(\lambda) = \text{sup}(A^\lambda) \tag{13}$$

Then this is compared to the base level, selecting the final best model \bar{A}_{max} as follows:

$$A_{max} = \text{max}\{\bar{A}_{nn}(\gamma), \bar{A}(\lambda)\} \tag{14}$$

This analysis is initially carried out for all the gene expression data available after selecting the top $q = 1\%$. In this case, the initial number of gene expression data per patient entails 566 rows of information. Then a Monte Carlo approach is followed, in which the number of rows is randomly reduced in each iteration by a random number β . This random number β is changed in each iteration and is strictly less than the total number of rows in the previous iteration. An example is summarized in Table 3. The rationale behind using a Monte Carlo simulation approach is that it is not feasible to estimate all the possible combinations of 566 genes, and hence some type of combinatorial approach needs to be used. This is a frequent situation in polygenic illness, such as IBD, in which a potentially large number of genes might be involved in the disease.

Table 3. Example of iterative algorithm testing different configurations of gene expressions.

Iteration	Initial N. Genes	β
0	566	30
1	536	125
2	411	58
3	353	9
4	344	215
\vdots	\vdots	\vdots

This process is repeated p times ($p = 100$), and the ten most accurate models are selected.

In the second section, a similar approach is followed but the mapping shown in Equation (1) has to be changed, as the objective is now to distinguish between ulcerative colitis and Crohn’s disease cases (the two major types of IBD). The mapping in this case is as follows (Equation (15)):

$$\forall j \in [1, u], \Phi_j = \{UC = 0, CD = 1\} \tag{15}$$

An alternative approach to the one presented is using a linear approach, such as, for instance, lasso regression [48,49]. Lasso regression offers the advantage that it makes some of the coefficients equal to zero, in practice reducing the number of inputs to the model. Using lasso regression, it is possible to reduce the number of genes selected for the classification model. In fact, lasso has become a frequently used feature selection algorithm [50,51].

3. Results

As previously described, the first step involves estimating a base level for the accuracy using artificial neural networks with simulations using 1 to 100 hidden layers. Each layer consists of 30 neurons. As it can be seen in Figure 4, increasing the number of layers does not necessarily translate into higher accuracy. The highest accuracy (testing dataset) obtained is 80.35% with a configuration including 920 hidden layers. The only other simulation reaching an accuracy above 80.00% is an ANN with 330 layers, reaching 80.10%. All the other simulations achieve a mean accuracy below 80.00%. No model has an accuracy below 70%. These results are obtained for a configuration of 74 rows (gene expression) which, as will be shown later, is the configuration that obtains the highest accuracy for the machine learning algorithm tested. As previously mentioned, the reported accuracy is the accuracy of the testing dataset, which is not used during the training phase.

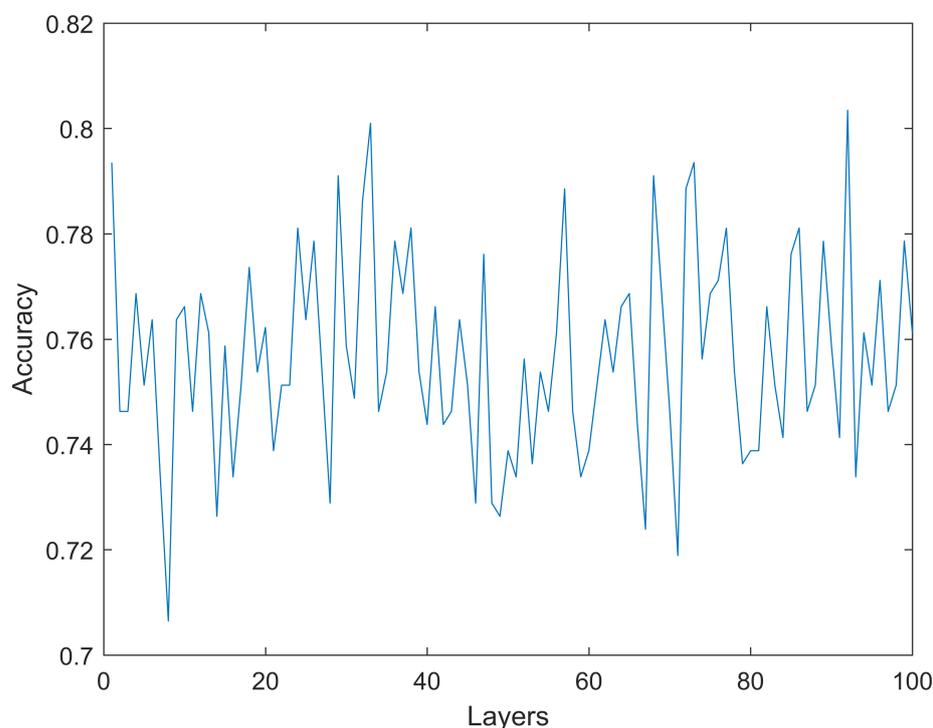


Figure 4. Accuracy of the neural network model for a range of number of artificial neurons. No model has an accuracy below 70% or higher than 80.35%.

Different machine learning algorithms are tested (as described in the Materials and Methods section). As an example, in Table 4, the accuracy results for one of the simulations are shown (140 gene expressions). In this specific case, the highest accuracy obtained is 81.5%. This accuracy is obtained by five different algorithms (Linear SVM, Fine Gaussian SVM, Medium Gaussian SVM, Coarse Gaussian SVM and Coarse KNN).

The results from the 10 most accurate simulations can be seen in Table 5. Of the ten most accurate results, nine use the bagged trees algorithm. The only other algorithm in the top ten most accurate models is the Subspace KNN. The highest accuracy is obtained for a model with 74 gene expression data, obtaining an accuracy, sensitivity and specificity of 84.2%, 82.6% and 84.4%, respectively. The list with these 74 genes can be found in Table 6.

Table 4. As an example, in this table, sample training with all 23 algorithms is shown. In this case, the model uses 140 gene expression data and the highest accuracy is 81.5%. This accuracy is actually reached by several algorithms (Linear SVM, Fine Gaussian SVM, Medium Gaussian SVM, Coarse Gaussian SVM and Coarse KNN).

Algorithm	Accuracy
Complex Tree	0.701
Medium Tree	0.783
Simple Tree	0.804
Linear Discriminant	0.645
Quadratic Discriminant	0.711
Logistic Regression	0.807
Linear SVM	0.815
Quadratic SVM	0.788
Cubic SVM	0.756
Fine Gaussian SVM	0.815
Medium Gaussian SVM	0.815
Coarse Gaussian SVM	0.815
Fine KNN	0.719
Medium KNN	0.770
Coarse KNN	0.815
Cosine KNN	0.768
Cubic KNN	0.764
Weighted KNN	0.773
Boosted Trees	0.805
Bagged Trees	0.804
Subspace Discriminant	0.812
Subspace KNN	0.748
RUSBoosted Trees	0.606

Table 5. Top ten models obtained according to the accuracy metric.

N. Genes	Algorithm	Accuracy	Sensitivity	Specificity
74	Bagged Trees	0.842	0.826	0.844
38	Subspace KNN	0.842	0.755	0.859
18	Bagged Trees	0.839	0.787	0.847
139	Bagged Trees	0.836	0.755	0.850
220	Bagged Trees	0.834	0.758	0.847
266	Bagged Trees	0.833	0.740	0.850
26	Bagged Trees	0.833	0.821	0.834
16	Bagged Trees	0.833	0.879	0.828
17	Bagged Trees	0.831	0.738	0.848
104	Bagged Trees	0.830	0.750	0.843

The results, when differentiating UC and CD cases, are not as accurate as when differentiating between control cases and IBD cases. This is in line with the expectations, as we are differentiating between two types of the same illness. These results are shown in Table 7. The most accurate result is obtained when using 562 gene expression data and the bagged trees algorithm. The accuracy, sensitivity and specificity are 73.4%, 79.0% and 71.2%, respectively. The list with these 562 genes can be found in the Supplementary Material.

Table 6. List of 74 genes selected by the algorithm.

B2M	RPS3	CHP1	SLC35A3
MALAT1	MAN2B1	ETNK1	PDIA3
EEF1A1	NDRG1	SLC1A2	DDX3X
MUC2	AHCYL2	GHITM	WDR1
FABP6	RPS14	MGAT4A	KLF5
KRT20	MYO1D	CLDN7	TSC22D1
CA1	A2M	COPZ2	RPL35A
FLNB	ADH1C	APOC3	SCP2
PHGR1	DDX17	SAT1	MATR3
IGKV1-5	FOS	ACE	CD46
CKB	RPL7	CD2AP	HNRNPH1
FABP1	SLC44A1	PAPSS2	PRKDC
FABP2	FN1	PDCD4	RPL37
CLDN4	RPL18	HPGD	LUM
TSPAN3	TDP2	UGT2A3	HSPA9
CDHR2	RPS12	UQCRC1	KIAA1109
CLTC	SPINT2	ST6GALNAC6	MIM24
COL1A2	RPL10A	ARF1	
ENO1	NCOA4	PRKACB	

Table 7. Top ten models obtained according to the accuracy metric distinguishing UC and CD patients.

N. Genes	Algorithm	Accuracy	Sensitivity	Specificity
562	Bagged Trees	0.734	0.790	0.712
66	Bagged Trees	0.728	0.679	0.767
24	Bagged Trees	0.718	0.665	0.742
37	Bagged Trees	0.718	0.821	0.687
564	Bagged Trees	0.712	0.909	0.671
132	Bagged Trees	0.704	0.929	0.676
49	Bagged Trees	0.704	0.679	0.719
15	Bagged Trees	0.700	0.713	0.697
550	Bagged Trees	0.694	0.871	0.659
277	Bagged Trees	0.673	0.616	0.717

As previously mentioned, an alternative approach to the one proposed is using lasso regression as a tool for the selection of inputs. The lasso approach selects 470 genes with the goodness-of-fit metric shown in Table 8. The accuracy and specificity results obtained in this approach are similar to those obtained in the proposed approach in the previous section. However, the sensitivity results from the lasso approach seem to be lower.

Table 8. Top ten models obtained using the lasso approach (470 genes) according to the accuracy metric distinguishing between control and UC and CD patients.

Algorithm	Accuracy	Sensitivity	Specificity
Medium KNN	0.817	0.667	0.817
Bagged Trees	0.817	0.667	0.817
Weighted KNN	0.815	0.500	0.816
Cubic KNN	0.807	0.143	0.815
Simple Tree	0.804	0.231	0.816
Subspace Dis.	0.804	0.405	0.829
Linear Dis.	0.802	0.433	0.842
Cosine KNN	0.802	0.300	0.819
Medium Tree	0.797	0.200	0.817
Subspace KNN	0.786	0.313	0.826

The lasso approach is also used to distinguish between UC and CD patients. In this case, the lasso approach selects 430 genes. The table with the goodness-of-fit results in this

approach is shown below (Table 9). The results using the lasso approach to distinguish between UC and DC patients are not as accurate as in the previous section. In both cases, using lasso or the proposed approach, differentiating between UC and DC patients appears to be more challenging than differentiating between control health individuals and patients with UC/CD. The lasso approach does not appear to increase the goodness of fit of the classification forecasts compared to the approach followed in the previous section.

Table 9. Top ten models obtained using the lasso approach (430 genes) according to the accuracy metric distinguishing between UC and CD patients.

Algorithm	Accuracy	Sensitivity	Specificity
Subspace Dis.	0.584	0.611	0.523
Logistic Reg.	0.572	0.617	0.503
Medium KNN	0.568	0.580	0.493
Cubic KNN	0.562	0.578	0.474
Weighted KNN	0.560	0.584	0.478
Simple Tree	0.558	0.569	0.412
Bagged Trees	0.558	0.583	0.474
Boosted Trees	0.556	0.580	0.467
Cosine KNN	0.550	0.574	0.448
Fine KNN	0.538	0.595	0.463

4. Discussion

Machine learning techniques are used to identify a set of 74 genes, which can be used, with an average accuracy of 84.2%, to distinguish between control (healthy individuals) and patients with inflammatory bowel disease. The specificity and sensitivity of this model are also relatively high at 82.6% and 84.4%, respectively. The selection of these 74 genes is carried out following a Monte Carlo simulation approach. Given that some of the symptoms of inflammatory bowel disease are common in other illnesses, it might be interesting to have another objective diagnostic tool. It is also interesting to observe that among multiple machine learning techniques used in the cohort of patients analyzed, the bagged trees approach seems to consistently achieve a high level of accuracy, particularly when compared to other, arguably more sophisticated machine learning techniques, such as artificial neural networks. The analysis controls for age, gender and region of the biopsy. The proportion of female and male cases is balanced, with 1174 female patients and 1316 male patients. The average age in the cohort is 44.9 years, covering a wide age range (from 19 to 82 years old). The results of the artificial neural networks include an optimization of the hyperparameters with simulations ranging from 1 to 1000 hidden layers. It is also observed that simply increasing the number of layers in an artificial neural network does not necessarily translate into better accuracy. It is also possible to distinguish between the two main types of IBD—Crohn’s disease and ulcerative colitis—but in this case with a lower level of accuracy. The accuracy, using this approach is 73.4%. The accuracy, sensitivity and specificity reported are those of the testing dataset. As normal practice, the data are divided into training and testing datasets in an attempt to increase the reproducibility of the analysis. Approximately 20% of the total cases are included in the testing dataset. The relatively large number of genes obtained in the best model is in line with the prevalent view in the existing literature that the illness is polygenic.

There is a high degree of heterogeneity in inflammatory bowel disease, leading to varied severity and evolution of the illness. The existing literature, see, for instance, Yamamoto et al. [52] or Ahmad et al. [33], points towards a polygenic illness with a complex interaction with environmental factors. Our results are consistent with this polygenic description. In this context, it is important to generate algorithms that are able to differentiate among control and patients as well as between different types of inflammatory bowel disease, namely Crohn’s disease and ulcerative colitis. A promising area of future research is to apply this type of approach in order to target treatments in a

more personalized way. It seems reasonable that there could be genetic differences among patients that can have a substantial impact on the outcome of the suggested treatments. This is particularly important in the context of inflammatory bowel disease, given the heterogeneity of the responses to treatments by different patients.

Some of the genes identified by the proposed algorithm are cited in the existing literature on intestinal-related illnesses. B2M was mentioned by Krzystek-Korpaczka et al. [53] in the context of bowel inflammation. There are other papers, such as that of Bednarz-Misa et al. [54], discussing B2M in the context of bowel inflammation and cancer. Another gene identified by the algorithm is MALAT1, which is also mentioned in the existing literature. Li et al. [55] suggested that MALAT1 maintains intestinal mucosal homeostasis in Crohn's disease. The authors concluded that the downregulation of MALAT1 contributes to the pathogenesis of CD. EEF1A1 was identified in a dog study as being involved in inflammatory bowel disease and cancer by Sahoo et al. [56]. The role of MUC2 in protecting the integrity of the mucosa was mentioned by Huang et al. [57]. The authors mentioned that it is possible to induce colitis in mice by suppressing the MUC2 gene. Heimel et al. [58] found high levels of expression of FABP2 and FABP6 when analyzing alterations in intestinal fatty acid metabolism in IBD. CA1 was mentioned by Xie et al. [59] as playing a role in IBD. PHGR1 was identified by Camilleri et al. [60] as potentially increasing the risk of diverticular disease of the colon. FABP1 was identified as a biomarker for Crohn's disease by Dooley et al. [61]. COL1A2 was mentioned by Prados et al. [62] in murine models of IBD. ENO1 was mentioned by Shkoda et al. [63] for its role in IBD pathobiology. Another gene selected by the algorithm and mentioned in the literature as being related to IBD is NDRG1 [64]. Song et al. [65] showed that ADH1C is downregulated in UC. FN1 was suggested by Al-Numan [66] to be related to the early onset of IBD. SPINT2 plays a role in epithelial adhesion [17]. CLDN7 is associated with colitis according to several authors [67,68]. Darsigny et al. [69] found a link between APOC3 and chronic inflammation in mice resembling IBD. KLF5 was identified by Dong et al. [70] as one of the genes downregulated in IBD. Gorenjak et al. [71] linked HSPA9 with IBD.

One of the challenges, and possible limitations, of this type of analysis is the fact that it is impossible to estimate all possible combinations of genes, and hence it is necessary to use some sort of combinatorial approach, such as the Monte Carlo model used to select the genes. There is also no indication that gene expression and IBD are related by an underlying linear model. Given this assumption, using machine learning techniques, which are adept to modeling nonlinear systems, seems like a reasonable approach. Another factor to take into account is that, while the cohort of cases is not small, including 2490 cases, it can always be larger.

5. Conclusions

Following a machine learning approach, it was possible to identify a list of genes that appear to be related to inflammatory bowel disease. Given the complexity of this illness, which appears to be caused by a combination of polygenic factors as well as environmental factors, which could, in principle, interact in a non-linear way, the illness was analyzed using non-linear models, such as machine learning techniques. This approach was able to distinguish, using a small number of genes, between patients with IBD and control (healthy) patients as well as patients with the two major forms of IBD, which are Crohn's disease and ulcerative colitis. In other words, the machine learning algorithms are able to classify different types of gene expression signatures associated with IBD. It might be possible in the future, when more data become available, to be able to distinguish between different genetic signatures of the illness that might potentially help develop more personalized treatments. This is important for an illness as heterogeneous as IBD, for which patients follow different evolutions and might present different clinical manifestations.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/medicina59071218/s1>, Supplementary Material: The list of 562 genes.

Author Contributions: Methodology, G.A.P. and R.C.; software, G.A.P.; validation, G.A.P. and R.C.; formal analysis, G.A.P. and R.C.; investigation, G.A.P. and R.C.; resources, G.A.P. and R.C.; data curation, G.A.P. and R.C.; writing—original draft preparation, G.A.P.; writing—review and editing, G.A.P. and R.C.; visualization, G.A.P. and R.C.; supervision, G.A.P. and R.C.; project administration, G.A.P. and R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Spanish Ministerio de Ciencia, Innovación y Universidades (PGC2018-094852-B-C21), and Universitat Jaume I (UJI-B2019-43).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are accessible at: <https://www.ncbi.nlm.nih.gov/geo/> (accessed on 1 June 2023).

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

Abbreviations

The following abbreviations are used in this manuscript:

IBD	Inflammatory Bowel Disease (IBD)
UC	Ulcerative Colitis
CD	Crohn's Disease
GEO	Gene Expression Omnibus
KNN	k-Nearest Neighbors
SVM	Support Vector Machine
ANN	Artificial Neural Network

References

1. Frolkis, A.; Dieleman, L.A.; Barkema, H.W.; Panaccione, R.; Ghosh, S.; Fedorak, R.N.; Madsen, K.; Kaplan, G.G. Environment and the inflammatory bowel diseases. *Can. J. Gastroenterol.* **2013**, *27*, 18–24. [[CrossRef](#)] [[PubMed](#)]
2. Rogler, G.; Singh, A.; Kavanaugh, A.; Rubin, D.T. Extraintestinal manifestations of inflammatory bowel disease: Current concepts, treatment, and implications for disease management. *Gastroenterol.* **2021**, *161*, 1118–1132. [[CrossRef](#)] [[PubMed](#)]
3. Seyedian, S.S.; Nokhostin, F.; Malamir, M.D. A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease. *J. Med. Life* **2019**, *12*, 113. [[CrossRef](#)] [[PubMed](#)]
4. Zhang, Y.; Li, Y. Inflammatory bowel disease: Pathogenesis. *World J. Gastroenterol.* **2014**, *20*, 91. [[CrossRef](#)] [[PubMed](#)]
5. Pithadia, A.B.; Jain, S. Treatment of inflammatory bowel disease (IBD). *Pharmacol. Rep.* **2011**, *63*, 629–642. [[CrossRef](#)]
6. Kaplan, G.G.; Windsor, J.W. The four epidemiological stages in the global evolution of inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* **2021**, *18*, 56–66. [[CrossRef](#)]
7. Xie, J.; Itzkowitz, S.H. Cancer in inflammatory bowel disease. *World J. Gastroenterol.* **2008**, *14*, 378. [[CrossRef](#)]
8. Lewandowski, K.; Kaniewska, M.; Więcek, M.; Szwarc, P.; Panufnik, P.; Tulewicz-Marti, E.; Walicka, M.; Franek, E.; Rydzewska, G. Risk factors for osteoporosis among patients with inflammatory bowel disease—Do we already know everything? *Nutrients* **2023**, *15*, 1151. [[CrossRef](#)]
9. Xavier, R.J.; Podolsky, D.K. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **2007**, *448*, 427–434. [[CrossRef](#)]
10. Liu, T.-C.; Stappenbeck, T.S. Genetics and pathogenesis of inflammatory bowel disease. *Annu. Rev. Pathol. Mech. Dis.* **2016**, *11*, 127–148. [[CrossRef](#)]
11. Neurath, M.F. Cytokines in inflammatory bowel disease. *Nat. Rev. Immunol.* **2014**, *14*, 329–342. [[CrossRef](#)] [[PubMed](#)]
12. Kuang, R.; O'Keefe, S.J.D.; Ramos del Aguila de Rivers, C.; Koutroumpakis, F.; Binion, D.G. Is salt at fault? Dietary salt consumption and inflammatory bowel disease. *Inflammatory Bowel Diseases* **2023**, *29*, 140–150. [[CrossRef](#)] [[PubMed](#)]
13. Baumgart, D.C.; Carding, S.R. Inflammatory bowel disease: Cause and immunobiology. *Lancet* **2007**, *369*, 1627–1640. [[CrossRef](#)] [[PubMed](#)]
14. Shanahan, F. Inflammatory bowel disease: Immunodiagnostics, immunotherapeutics, and ecotherapeutics. *Gastroenterology* **2001**, *120*, 622–635. [[CrossRef](#)] [[PubMed](#)]
15. Tamboli, C.P.; Neut, C.; Desreumaux, P.; Colombel, J.F. Dysbiosis in inflammatory bowel disease. *Gut* **2004**, *53*, 1–4. [[CrossRef](#)] [[PubMed](#)]
16. Chang, J.T. Pathophysiology of inflammatory bowel diseases. *New Engl. J. Med.* **2020**, *383*, 2652–2664. [[CrossRef](#)] [[PubMed](#)]
17. Graham, D.B.; Xavier, R.J. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature* **2020**, *578*, 527–539. [[CrossRef](#)]

18. Strober, W.; Fuss, I.; Mannon, P. The fundamental basis of inflammatory bowel disease. *J. Clin. Investig.* **2007**, *117*, 514–521. [[CrossRef](#)]
19. Loddo, I.; Romano, C. Inflammatory bowel disease: Genetics, epigenetics, and pathogenesis. *Front. Immunol.* **2015**, *6*, 551. [[CrossRef](#)]
20. Khor, B.; Gardet, A.; Xavier, R.J. Genetics and pathogenesis of inflammatory bowel disease. *Nature* **2011**, *474*, 307–317. [[CrossRef](#)]
21. Duerr, R.H. Update on the genetics of inflammatory bowel disease. *J. Clin. Gastroenterol.* **2003**, *37*, 358–367. [[CrossRef](#)] [[PubMed](#)]
22. McGovern, D.P.B.; Kugathasan, S.; Cho, J.H. Genetics of inflammatory bowel diseases. *Gastroenterology* **2015**, *149*, 1163–1176. [[CrossRef](#)] [[PubMed](#)]
23. Eguchi, R.; Karim, M.B.; Hu, P.; Sato, T.; Ono, N.; Kanaya, S.; Altaf-Ul-Amin, M. An integrative network-based approach to identify novel disease genes and pathways: A case study in the context of inflammatory bowel disease. *BMC Bioinform.* **2018**, *19*, 1–12. [[CrossRef](#)] [[PubMed](#)]
24. Lees, C.W.; Satsangi, J. Genetics of inflammatory bowel disease: Implications for disease pathogenesis and natural history. *Expert Rev. Gastroenterol. Hepatol.* **2009**, *3*, 513–534. [[CrossRef](#)]
25. Cho, J.H.; Abraham, C. Inflammatory bowel disease genetics: Nod2. *Annu. Rev. Med.* **2007**, *58*, 401–416. [[CrossRef](#)]
26. Bonen, D.K.; Cho, J.H. The genetics of inflammatory bowel disease. *Gastroenterology* **2003**, *124*, 521–536. [[CrossRef](#)]
27. Kakuta, Y.; Naito, T.; Kinouchi, Y.; Masamune, A. Current Status and Future Prospects of Inflammatory Bowel Disease Genetics. *Digestion* **2023**, *104*, 7–15. [[CrossRef](#)]
28. Mathew, C.G.; Lewis, C.M. Genetics of inflammatory bowel disease: Progress and prospects. *Digestion* **2004**, *13*, 161–168. [[CrossRef](#)]
29. Achkar, J.; Duerr, R. The expanding universe of inflammatory bowel disease genetics. *Curr. Opin. Gastroenterol.* **2008**, *24*, 429–434. [[CrossRef](#)]
30. Van Limbergen, J.; Russell, R.K.; Nimmo, E.R.; Satsangi, J. The genetics of inflammatory bowel disease. *Off. J. Am. Coll. Gastroenterol.* **2007**, *102*, 2820–2831. [[CrossRef](#)]
31. Stoll, M.; Corneliussen, B.; Costello, C.M.; Waetzig, G.H.; Mellgard, B.; Koch, W.A.; Rosenstiel, P.; Albrecht, M.; Croucher, P.J.P.; Seegert, D. Genetic variation in DLG5 is associated with inflammatory bowel disease. *Nat. Genet.* **2004**, *36*, 476–480. [[CrossRef](#)]
32. Cleynen, I.; Boucher, G.; Jostins, L.; Schumm, L.P.; Zeissig, S.; Ahmad, T.; Andersen, V.; Andrews, J.M.; Annesse, V.; Brand, S. Inherited determinants of Crohn’s disease and ulcerative colitis phenotypes: A genetic association study. *Lancet* **2016**, *387*, 156–167. [[CrossRef](#)] [[PubMed](#)]
33. Ahmad, T.; Satsangi, J.; McGovern, D.; Bunce, M.; Jewell, D.P. The genetics of inflammatory bowel disease. *Aliment. Pharmacol. Ther.* **2001**, *15*, 731–748. [[CrossRef](#)] [[PubMed](#)]
34. Hazel, K.; O’Connor, A. Emerging treatments for inflammatory bowel disease. *Ther. Adv. Chronic Dis.* **2020**, *11*, 1–12. [[CrossRef](#)] [[PubMed](#)]
35. Dhillon, A.; Singh, A.; Bhalla, V.K. A systematic review on biomarker identification for cancer diagnosis and prognosis in multi-omics: From computational needs to machine learning and deep learning. *Arch. Comput. Methods Eng.* **2023**, *30*, 917–949. [[CrossRef](#)]
36. Bhatt, M.; Shende, P. Advancement in Machine Learning: A Strategic Lookout from Cancer Identification to Treatment. *Arch. Comput. Methods Eng.* **2023**, *2023*, 1–16. [[CrossRef](#)]
37. Alfonso Perez, G.; Castillo, R. Identification of Systemic Sclerosis through Machine Learning Algorithms and Gene Expression. *Mathematics* **2022**, *10*, 4632. [[CrossRef](#)]
38. Mossotto, E.; Ashton, J.J.; Coelho, T.; Beattie, R.M.; MacArthur, B.D.; Ennis, S.J.S.R. Classification of paediatric inflammatory bowel disease using machine learning. *Sci. Rep.* **2017**, *7*, 1–10. [[CrossRef](#)]
39. Kohli, A.; Holzwanger, E.A.; Levy, A.N. Emerging use of artificial intelligence in inflammatory bowel disease. *World J. Gastroenterol.* **2020**, *26*, 6923. [[CrossRef](#)]
40. Gardiner, L.; Carrieri, A.P.; Bingham, K.; Macluskie, G.; Bunton, D.; McNeil, M.; Pyzer-Knapp, E.O. Combining explainable machine learning, demographic and multi-omic data to inform precision medicine strategies for inflammatory bowel disease. *PLoS ONE* **2022**, *17*, e0263248. [[CrossRef](#)]
41. 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. [[CrossRef](#)] [[PubMed](#)]
42. Wei, Z.; Wang, W.; Bradfield, J.; Li, J.; Cardinale, C.; Frackelton, E.; Kim, C.; Mentch, F.; Van Steen, K.; Visscher, P.M. 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](#)] [[PubMed](#)]
43. 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](#)] [[PubMed](#)]
44. Coelho, T.; Mossotto, E.; Gao, Y.; Haggarty, R.; Ashton, J.J.;Batra, A.; Stafford, I.S.; Beattie, R.M.; Williams, A.P.; Ennis, S. Immunological profiling of paediatric inflammatory bowel disease using unsupervised machine learning. *J. Pediatr. Gastroenterol. Nutr.* **2020**, *70*, 833–840. [[CrossRef](#)] [[PubMed](#)]
45. Smolander, J.; Dehmer, M.; Emmert-Streib, F. Comparing deep belief networks with support vector machines for classifying gene expression data from complex disorders. *FEBS Open Bio.* **2019**, *9*, 1232–1248. [[CrossRef](#)]

46. 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* **2017**, *12*, 1438. [[CrossRef](#)]
47. Argmann, C.; Hou, R.; Ungaro, R.C.; Irizar, H.; Al-Taie, Z.; Huang, R.; Kosoy, R.; Venkat, S.; Song, W.; Di'Narzo, A.F. Biopsy and blood-based biomarker of inflammation in IBD. *Gut* **2022**, *2022*, 1271–1287. [[CrossRef](#)]
48. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser.* **1996**, *58*, 267–288. [[CrossRef](#)]
49. Roth, V. The generalized LASSO. *IEEE Trans. Neural Netw.* **2004**, *15*, 16–28. [[CrossRef](#)]
50. Muthukrishnan, R.; Rohini, R. LASSO: A feature selection technique in predictive modeling for machine learning. In Proceedings of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 24 October 2016; pp. 18–20.
51. Wu, F.; Yuan, Y.; Zhuang, Y. Heterogeneous feature selection by group lasso with logistic regression. In Proceedings of the 18th ACM international conference on Multimedia, San Jose, CA, USA, 3–5 June 2010; pp. 983–986.
52. Yamamoto-Furusho, J.K. Genetic factors associated with the development of inflammatory bowel disease. *J. Gastroenterol.* **2007**, *13*, 5594. [[CrossRef](#)]
53. Krzystek-Korpacka, M.; Diakowska, D.; Bania, J.; Gamian, A. Expression stability of common housekeeping genes is differently affected by bowel inflammation and cancer: Implications for finding suitable normalizers for inflammatory bowel disease studies. *Inflamm. Bowel Dis.* **2014**, *20*, 1147–1156. [[CrossRef](#)] [[PubMed](#)]
54. Bednarz-Misa, I.; Neubauer, K.; Zacharska, E.; Kapturkiewicz, B.; Krzystek-Korpacka, M. Whole blood ACTB, B2M and GAPDH expression reflects activity of inflammatory bowel disease, advancement of colorectal cancer, and correlates with circulating inflammatory and angiogenic factors: Relevance for real-time quantitative PCR. *Adv. Clin. Exp. Med.* **2020**, *29*. [[CrossRef](#)] [[PubMed](#)]
55. Li, Y.; Zhu, L.; Chen, P.; Wang, Y.; Yang, G.; Zhou, G.; Li, L.; Feng, R.; Qiu, Y.; Han, J. MALAT1 maintains the intestinal mucosal homeostasis in Crohn's disease via the miR-146b-5p-CLDN11/NUMB pathway. *J. Crohn's Colitis* **2021**, *15*, 1542–1557. [[CrossRef](#)] [[PubMed](#)]
56. Sahoo, D.K.; Borchering, D.C.; Chandra, L.; Jergens, A.E.; Atherly, T.; Bourgois-Mochel, A.; Ellinwood, N.M.; Snella, E.; Severin, A.J.; Martin, M. Differential transcriptomic profiles following stimulation with lipopolysaccharide in intestinal organoids from dogs with inflammatory bowel disease and intestinal mast cell tumor. *Cancers* **2021**, *14*, 3525. [[CrossRef](#)] [[PubMed](#)]
57. Huang, E.Y.; Inoue, T.; Leone, V.A.; Dalal, S.; Touw, K.; Wang, Y.; Musch, M.W.; Theriault, B.; Higuchi, K.; Donovan, S. Using corticosteroids to reshape the gut microbiome: Implications for inflammatory bowel diseases. *Inflamm. Bowel Dis.* **2015**, *21*, 963–972. [[CrossRef](#)] [[PubMed](#)]
58. Heimerl, S.; Moehle, C.; Zahn, A.; Boettcher, A.; Stremmel, W.; Langmann, T.; Schmitz, G. Alterations in intestinal fatty acid metabolism in inflammatory bowel disease. *Biochim. Biophys. Acta Mol. Basis Dis.* **2006**, *1762*, 341–350. [[CrossRef](#)]
59. Xie, D.; Zhang, Y.; Qu, H. Crucial genes of inflammatory bowel diseases explored by gene expression profiling analysis. *Scand. J. Gastroenterol.* **2018**, *53*, 685–691. [[CrossRef](#)]
60. Camilleri, M.; Sandler, R.S.; Peery, A.F. Etiopathogenetic mechanisms in diverticular disease of the colon. *Cell. Mol. Gastroenterol. Hepatol.* **2020**, *9*, 15–32. [[CrossRef](#)]
61. Dooley, T.P.; Curto, E.V.; Reddy, S.P.; Davis, R.L.; Lambert, G.W.; Wilborn, T.W.; Elson, C.O. Regulation of gene expression in inflammatory bowel disease and correlation with IBD drugs. Screening by DNA microarrays. *Inflamm. Bowel Dis.* **2004**, *10*, 1–14. [[CrossRef](#)]
62. Prados, M.E.; García-Martín, A.; Unciti-Broceta, J.D.; Palomares, B.; Collado, J.A.; Minassi, A.; Calzado, M.A.; Appendino, G.; Muñoz, E. Betulinic acid hydroxamate prevents colonic inflammation and fibrosis in murine models of inflammatory bowel disease. *Acta Pharmacol. Sin.* **2021**, *42*, 1124–1138. [[CrossRef](#)]
63. Shkoda, A.; Werner, T.; Daniel, H.; Gunckel, M.; Rogler, G.; Haller, D. Differential protein expression profile in the intestinal epithelium from patients with inflammatory bowel disease. *J. Proteome Res.* **2007**, *6*, 1114–1125. [[CrossRef](#)] [[PubMed](#)]
64. Knyazev, E.; Maltseva, D.; Raygorodskaya, M.; Shkurnikov, M. HIF-dependent NFATC1 activation upregulates ITGA5 and PLAUR in intestinal epithelium in inflammatory bowel disease. *Front. Genet.* **2021**, *2277*, 1–12. [[CrossRef](#)] [[PubMed](#)]
65. Song, F.; Zhang, Y.; Pan, Z.; Hu, X.; Zhang, Q.; Huang, F.; Ye, X.; Huang, P. The role of alcohol dehydrogenase 1C in regulating inflammatory responses in ulcerative colitis. *Front. Genet.* **2021**, *192*, 114691. [[CrossRef](#)] [[PubMed](#)]
66. Al-Numan, H.H.; Jan, R.M.; Al-Saud, N.B.S.; Rashidi, O.M.; Alrayes, N.M.; Alsufyani, H.A.; Mujalli, A.; Shaik, N.A.; Mosli, M.H.; Elango, R. Exome sequencing identifies the extremely rare ITGAV and FN1 variants in early onset inflammatory bowel disease patients. *Front. Pediatr.* **2022**, *10*, 859074. [[CrossRef](#)]
67. Ding, Y.; Wang, K.; Xu, C.; Hao, M.; Li, H.; Ding, L. Intestinal Claudin-7 deficiency impacts the intestinal microbiota in mice with colitis. *BMC Gastroenterol.* **2022**, *22*, 24. [[CrossRef](#)]
68. Wang, K.; Ding, Y.; Xu, C.; Hao, M.; Li, H.; Ding, L. Cldn-7 deficiency promotes experimental colitis and associated carcinogenesis by regulating intestinal epithelial integrity. *Oncoimmunology* **2021**, *10*, 1923910. [[CrossRef](#)]
69. Darsigny, M.; Babeu, J.P.; Dupuis, A.A.; Furth, E.; Seidman, E.G.; Lévy, E.; Verdu, E.F.; Gendron, F.P.; Boudreau, F. Loss of hepatocyte-nuclear-factor-4 α affects colonic ion transport and causes chronic inflammation resembling inflammatory bowel disease in mice. *PLoS ONE* **2009**, *4*, 7609. [[CrossRef](#)]

70. Dong, J.T.; Chen, C. Essential role of KLF5 transcription factor in cell proliferation and differentiation and its implications for human diseases. *Cell. Mol. Life Sci.* **2009**, *66*, 2691–2706. [[CrossRef](#)]
71. Gorenjak, M.; Jezernik, G.; Krušič, M.; Skok, P.; Potočnik, U. Identification of Novel Loci Involved in Adalimumab Response in Crohn's Disease Patients Using Integration of Genome Profiling and Isoform-Level Immune-Cell Deconvoluted Transcriptome Profiling of Colon Tissue. *Pharmaceutics* **2022**, *14*, 1893. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.