Neural Computing and Applications (2021) 33:14133–14146
https://doi.org/10.1007/s00521-021-06055-x

ORIGINAL ARTICLE

A new approach to predict ulcerative colitis activity through standard clinical–biological parameters using a robust neural network model

Iolanda V. Popa[1,2] · Alexandru Burlacu[1,3] · Otilia Gavrilescu[1,2] · Mihaela Dranga[1,2] · Cristina Cijevschi Prelipcean[1,2] · Cătălina Mihai[1,2]

Received: 28 January 2020 / Accepted: 19 April 2021 / Published online: 2 May 2021

Abstract

Colonoscopy is the "gold" standard for evaluating disease activity in ulcerative colitis (UC). An important area of research is finding a cost-efficient, non-invasive solution for estimating disease activity. We aimed to develop and validate a neural network (NN) model that uses routinely available clinical–biological variables to predict UC activity. Standard clinical–biological parameters and endoscopic Mayo score from 386 UC patient records were collected. A training set (n = 285), a test set (n = 71) and a validation set (n=30) were used for constructing and validating three NN models. The first two models predicted the active/inactive endoscopic disease status through a binary output. The third model estimated the complete endoscopic Mayo score through a categorical output. First model (with seven categorical and 13 continuous input variables) obtained an accuracy of 94.37% on the test set and 93.33% on the validation set. The second model (with 12 biological input parameters) achieved an accuracy of 88.73% on the test set and 83.33% on the validation set. The third model used the same input variables as the first model obtaining an accuracy of 76.06% on the test set and 80% on the validation set. We designed an accurate and non-invasive artificial intelligence solution to estimate disease activity, other than colonoscopy. Our NN model achieved better results than pooled performance metrics of fecal calprotectin (the best non-invasive marker to date) investigated in UC. Given these promising results, we envision introducing of a non-invasive algorithm for routinely predicting disease activity shortly.

Keywords Neural networks · Predictive model · Ulcerative colitis · Disease activity assessment · Machine learning

1 Introduction

Ulcerative colitis (UC) is a type of inflammatory bowel disease (IBD) characterized by recurrent episodes of colorectal inflammation resulting in clinical relapses and remissions. Disease flares influence UC patients' quality of life and productivity, even in mild cases [1], whereas more severe disease relapses can be debilitating with serious, seldom life-threatening, complications [2]. Even patients with sub-clinical disease activity (clinical remission with endoscopic or histopathologic active disease) experience extraintestinal manifestations and discomfort [1]. From the mildest to the most severe, all disease forms have a negative socio-economic impact [3] that could be reduced by a more rigorous disease activity monitoring to indicate early signs of sub-clinical relapse emergence.

The "gold" standard for the evaluation of disease activity in UC is colonoscopy [4]. The European guidelines traditionally use endoscopic disease activity as a therapeutic target and in the process of clinical decision-making [4–6]. However, colonoscopy has significant drawbacks since its invasiveness yields inherent risks (perforation, lower gastrointestinal bleeding, cardiovascular and cerebrovascular events) [7].

Modern computational strategies in artificial intelligence (AI)/machine learning (ML) offer reliable alternatives to conventional diagnostic methods in the last years. To date, several papers describing ML models have addressed specific challenges of IBD regarding gastrointestinal image

✉ Alexandru Burlacu
alexandru.burlacu@umfiasi.ro

1 University of Medicine and Pharmacy "Gr. T. Popa", Iasi, Romania

2 Institute of Gastroenterology and Hepatology, Iasi, Romania

3 Department of Interventional Cardiology, Cardiovascular Diseases Institute, Iasi, Romania

analysis [8, 9], phenotype prediction [10], disease course prediction after treatment [11–13] and disease subtype classification [14]. However, no study was designated for the non-invasive estimation of UC endoscopic activity.

High-performance ML approaches for predicting disease severity based on electronic health records have been previously described for asthma [15], congestive heart failure [16] and sepsis [17]. The first non-invasive ML solution for predicting endoscopic disease severity in UC based on standard clinical parameters could be outstanding.

Therefore, our goal is to develop a neural network (NN) model to predict the disease activity in UC based on routinely available clinical variables, evaluate the model's predictive power and validate it on an independent patient cohort, laying the foundations for future use in clinical practice.

2 Materials and methods

2.1 Study design and participants

An observational retrospective single-center cohort study was conducted on a sample of 386 UC patient records. All patients were admitted to the Institute of Gastroenterology and Hepatology, "Sf. Spiridon" Hospital Iasi–Romania, between March 2011 and October 2019. Pre-diagnosed and newly diagnosed confirmed UC patients who underwent a colonoscopy for disease assessment were included. Patients were excluded if they were in evidence with concurrent disorders (infections, autoimmune and inflammatory conditions, cirrhosis, neoplasia, hemodialysis) capable of influencing medical parameters.

All patients provided written informed consent. The study has full ethical approval from the Research Ethics Commission of the "Gr. T. Popa" University of Medicine and Pharmacy Iasi—Romania and "Sf. Spiridon" Hospital Iasi—Romania Ethics Committee. No sex-based or racial/ethnic-based differences were present.

2.2 Clinical protocol

Pre-diagnosed UC patients were admitted for treatment monitoring or disease worsening. Newly UC diagnosed cases were admitted for typical or atypical onset of digestive symptoms, including rectal bleeding, diarrhea, abdominal pain, urgency and incontinence. According to the European consensus guidelines, a "gold standard" for UC diagnosis is established by clinical, biological, imaging, endoscopic and histopathologic findings [6]. Patients underwent a medical history interview, physical examination, routine laboratory tests and colonoscopy with biopsy to diagnose or assess already diagnosed UC disease. Patients were investigated following the European standard protocols. Only patients with a confirmed diagnosis of UC were included.

2.3 Data collection

Documented **clinical parameters** were: age, gender, smoking status, number of stools/day and presence of diarrhea, tenesmus, lower gastrointestinal bleeding (LGB), abdominal pain, weight loss, asthenia and pallor. Smoking status was a categorical variable with three possible values: 0—smoker, 1—non-smoker and 2—former smoker. Several stools/day was represented as a continuous variable. The presence of diarrhea, tenesmus, LGB, abdominal pain, weight loss, asthenia and pallor were represented as binary categorical variables (1 indicating presence and 0—absence of referred symptom).

Laboratory parameters documented were: red blood cells (RBC), white blood cells (WBC), platelets (PLT), hemoglobin (HGB), hematocrit (HCT), plateletcrit (PCT), platelet distribution width (PDW), mean platelet volume (MPV), platelet large cell ratio (PLCR), neutrophils (NEUT), lymphocytes, monocytes (MONO), C reactive protein (CRP), erythrocyte sedimentation rate/1h (ESR), fibrinogen, serum iron (SI), ferritin, total proteins (TP), albumin, alpha 1 globulins (A1G), alpha 2 globulins, beta 1 globulins, beta 2 globulins, gamma globulins, glucose.

Colonoscopy was performed on the EVIS EXERA II endoscopy system (Olympus America). The procedures were carried out by specialist physicians from the Gastroenterology and Hepatology Institute, Iasi—Romania. According to the endoscopic Mayo score, the colonoscopic findings were represented as a categorical variable with four possible values (from 0 to 3), as recommended by the European consensus guidelines [4, 6]. A patient was considered to have endoscopic remission if the Mayo score was 0 or 1. Similarly, active disease was considered for Mayo score 2 or 3. Subsequent colonoscopic examinations in an interval higher than one month were recorded separately.

2.4 Management of missing values

Documented continuous variables (biological parameters and number of stools/day) were standardized in the range [0, 1]. Missing values were assigned using multivariate imputation by chained equations method implemented by the MICE package in R Studio Version 1.2.1335 © 2009–2019 RStudio, Inc. Build 1379 (f1ac3452). Missing continuous variables were assigned by applying the Bayesian regression built-in method, while categorical data were imputed using the logistic regression built-in method.

2.5 Standard statistics for feature selection

To use the ANOVA test, the fulfillment of the specific assumptions was checked using R Studio. We verified that the group samples were drawn from normally distributed populations using the ggqqplot function (ggpubr package) for the univariate analysis. The mqqnorm function (RVAideMemoire package) was used to test multivariate normality. The ggqqplot and mqqnorm functions were preferred to the univariate and multivariate Shapiro–Wilk test since the population size is greater than 50. The homogeneity of variances was assessed using the bartlett.test function considering the Mayo score as the grouping variable. The outliers were detected with the help of the identify_outliers function (rstatix package). The ANOVA prerequisites were assessed with and without the outliers, and the results were compared.

ANOVA with Holm adjustments in R Studio was used to determine whether significant differences between the four Mayo groups existed for each continuous parameter. Statistical significance was considered for $p \leq .05$. Only parameters with significant differences between at least four pairs of groups were included in further analysis. If any two of the selected continuous variables had high intercorrelation with a Pearson coefficient ≥ 0.9, one of them was removed.

Chi-square test of independence was performed to examine whether there is a relationship between each categorical parameter and the Mayo score. The significance level was considered at .01. Variables for which the null hypothesis was rejected (those proved to have an association with Mayo score) were selected for further investigation.

2.6 Neural network models: Construction and evaluation

Initial data (356 patient records) were randomly divided into a training set of 285 records (80%) and a test set of 71 records (20%) such that variables distributions in each set were similar to those in the original dataset. Other 30 patient records from the same medical center were added independently to be used as a validation set. Mayo categories were not equally represented in the train and test sets, while the validation set had a balanced distribution of Mayo classes.

Three multilayered perceptron classifiers were developed based on the training set.

Classifiers were constructed using the mlpML method within the caret::train function in R Studio. A 10-fold cross-validation was used to reduce the problem of overfitting [18]. The 10-fold cross-validation was repeated ten times to reduce the error in the estimate of mean model performance. Synthetic minority over-sampling technique (SMOTE) was used with caret::train function to overcome the issue of imbalanced data [19]. Several activation functions were evaluated in terms of model's performance: Hyperbolic tangent (TanH), Softmax, Signum, Sinus, Elliott, Threshold and Gaussian. The function contributing to the highest performance metrics was selected as the transfer function of the hidden and output neurons. The caret::train function automatically tuned three hyperparameters for the mlpML method. The automatically tuned parameters were the number of neurons in each of the three hidden layers corresponding to the mlpML method design. The initialization function used was randomized weights. Standard backpropagation was employed as the learning function with a 0.2 step width of the gradient descent. Topological order was used as the update function. All parameters are tuned to maximize the model's performance.

The first two classifiers were used to predict whether a UC patient has endoscopic activity or remission based on all 20 parameters chosen by the feature selection method (first classifier) or based only on biological parameters (second classifier). The second classifier was built as it is of interest to construct and evaluate a model based only on objective data. The third classifier was used for the prediction of the Mayo score based on all 20 parameters. The first two classifiers have a binary output, while the third has a categorical output with four possible values (Mayo score from 0 to 3).

A detailed implementation flowchart of the three proposed classifiers is illustrated in Figure 1.

The developed NNs were evaluated on the test set and validation based on accuracy (ACC) of classification. Where applicable, area under the receiver operating characteristic curve (AUC), sensitivity (SE), specificity (SP), positive and negative predictive values (PPV and NPV) were also determined.

To further evaluate the proposed classifiers, we also conducted several comparisons with other algorithms concerning the training time and classification accuracy. Our models were compared to three methods implemented by the caret::train function: random forest (rf), support vector machine with linear kernel (svmLinear) and support vector machine with radial basis function kernel (svmRadial).

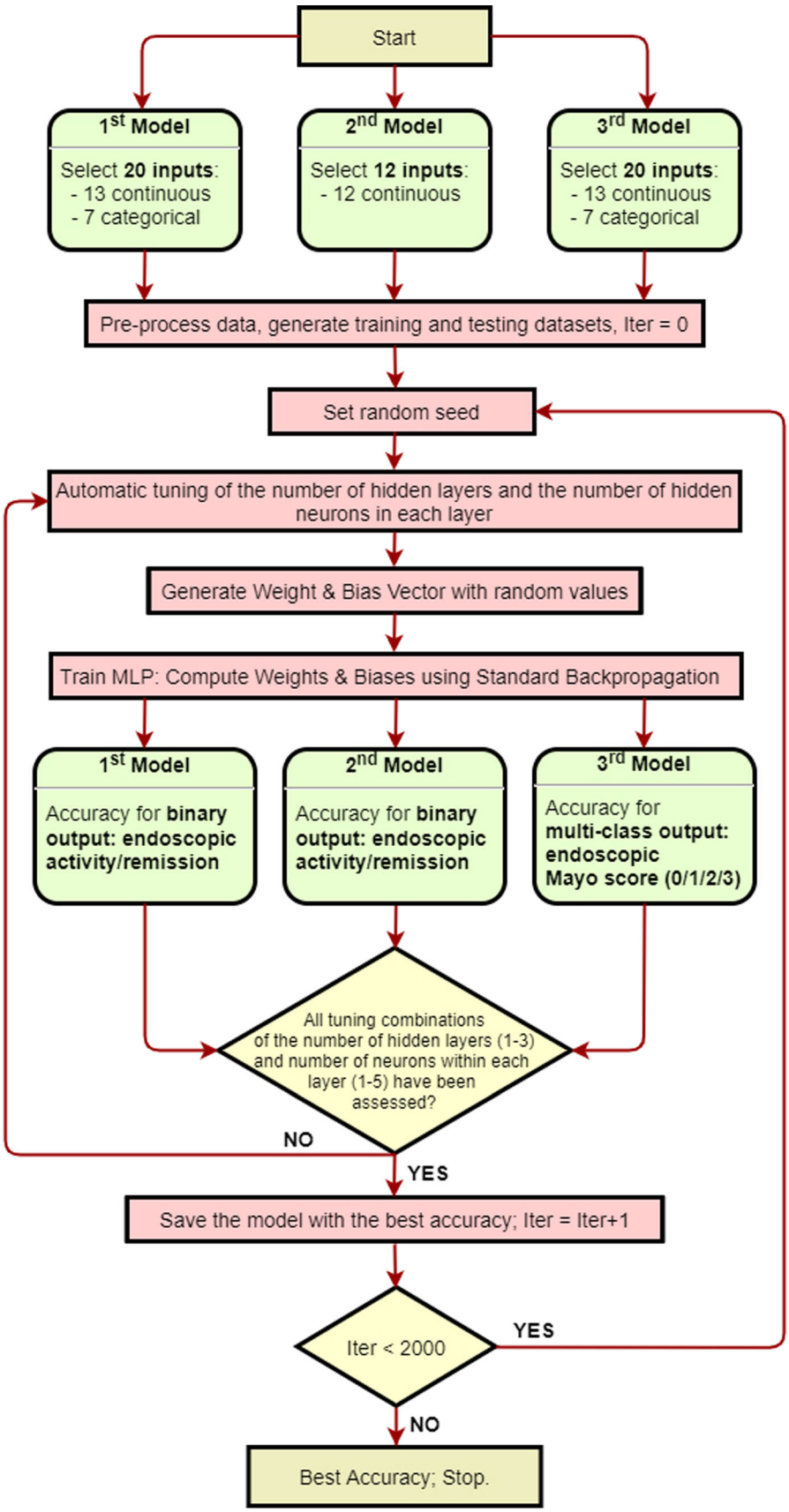

Figure 1 Legend. Detailed implementation flowchart of the three multilayered perceptrons.

3 Results

3.1 Patient characteristics

Of all 386 patient records, 257 (66%) were males and 129 (34%) females. The age range of the participants was 18-82. The distribution of the Mayo groups was imbalanced, with the Mayo 2 group containing three times more records than each of the other groups: 66 records were classified with an endoscopic Mayo score of 0, 65 with a Mayo score of 1, 189 had a moderate endoscopic activity (Mayo 2), and 66 had a severe disease activity (Mayo 3).

Selected clinical characteristics and laboratory findings for all patient records, and each of the Mayo classes are summarized in Table 1. The means for each featured variable tend to increase or decrease with the Mayo score.

3.2 Management of missing values

A total of 367 (4,75%) missing values were imputed using the MICE package as follows: tenesmus—3, LGB—1, abdominal pain—2, weight loss—4, asthenia—2, palor—4, MONO—1, ESR—55, fibrinogen—90, SI—26, TP—61, A1G—78, CRP—33, PDW—7.

3.3 Standard statistics for feature selection

Our data met the ANOVA assumptions. Populations are normally distributed and have common variances, provided that the Mayo score is the grouping variable. Moreover, the multivariate normality was confirmed by the multi-normal Q-Q plot resulted after applying the function mqqnorm in R. The data have a normal multivariate distribution as the points in the multi-normal Q-Q plot tend to lie on a straight diagonal line (Figure S1). The presence of the outliers did not affect the results of the analysis.

Pairwise comparisons were carried out using ANOVA with Holm adjustments to select 18 continuous variables with significant differences between at least four of the six Mayo groups comparisons. Significant differences between all six Mayo classes were found for one parameter (number of stools/day). There were significant differences between five group comparisons concerning ten variables (WBC, PLT, MONO, NEUT, PCT, ESR, fibrinogen, SI, TP, A1G) and between four group comparisons concerning seven variables (CRP, RBC, HGB, HCT, PDW, MPV, PLCR). The results of the ANOVA pairwise comparisons are given in Table 2.

The next step was to identify and reduce selected features that are highly intercorrelated. Figure 2 shows the

Table 1 Clinical and biological parameters for all patient records and each Mayo group.

	All	Mayo 0	Mayo 1	Mayo 2	Mayo 3
Number of records	386	66	65	189	66
Gender (male:female)	257:129	43:23	38:27	122:67	54:12
Age (years)	44.8 ± 13.9	43.6 ± 12.8	44.6 ± 12.1	44.4 ± 14.4	122 ± 14.9
Number of stools/day	4.9 ± 3.8	1.3 ± 0.8	2.9 ± 2.27	5.6 ± 3.4	8.8 ± 4
LGB	257 (66.6%)	1 (1.5%)	24 (36.9%)	168 (88.9%)	64 (97%)
Diarrhea	265 (68.7%)	6 (9.1%)	24 (36.9%)	169 (89.4%)	64 (97%)
Tenesmus	184 (47.7%)	0	12 (18.5%)	121 (64%)	51 (77.3%)
Abdominal pain	226 (58.5%)	10 (15.2%)	23 (35.4%)	140 (74.1%)	53 (80.3%)
Weight loss	118 (38.6%)	2 (3%)	7 (10.8%)	70 (37%)	39 (59.1%)
Asthenia	210 (54.4%)	7 (10.6%)	23 (35.4%)	133 (70.4%)	47 (71.2%)
Pallor	97 (25.1)	1 (1.5%)	8 (12.3%)	59 (31.2%)	27 (40.9%)
WBC $*10^3/\mu L$	8.2 ± 3.2	6.6 ± 1.6	7.4 ± 2.3	8.4 ± 3.4	10.1 ± 3.8
HGB g/dL	13.3 ± 2.1	14.3 ± 1.5	13.7 ± 1.6	13.2 ± 2	12.4 ± 2.6
RBC $*100^3/\mu L$	4.7 ± 0.5	4.9 ± 0.4	4.8 ± 0.4	4.7 ± 0.5	4.4 ± 0.7
PLT $*10^3/\mu L$	311.2 ± 111.6	252 ± 52.1	269.2±95.7	318.1±87.2	392.3±167.1
MONO $*10^3/\mu L$	0.68 ± 0.34	0.52 ± 0.15	0.56 ± 0.17	0.68 ± 0.32	1 ± 0.45
PDW fl	12.3 ± 2.2	13.1 ± 1.6	13.2 ±2.9	12.1 ± 2	11.7 ± 1.9
ESR mm/1h	15.83 ± 19.6	5.3 ± 5.7	9.1 ± 11.6	16.3 ± 15.4	32.9 ± 32
Fibrinogen mg/dl	388.1±83	338.1±74.2	359.5±56.3	394.7±68.7	465.1±100.2
CRP mg/dl	1.61 ± 3.4	0.3 ± 0.4	0.4 ± 0.4	1.3 ± 2.3	5.3 ± 6.1
SI µg/dl	67.3 ± 40.8	88.6 ± 34.3	82.6 ± 41	62.3 ± 40.5	41.6 ± 28.4
TP g/dl	7.4 ± 0.7	7.68 ± 0.49	7.65 ± 0.58	7.36 ± 0.67	6.89 ± 0.85
A1G %	2.8 ± 1.2	1.9 ± 0.3	2.3 ± 0.5	2.8 ± 0.9	4.2 ± 1.8

Table 2 ANOVA pairwise comparisons concerning continuous parameters.

	P value for Mayo classes pairwise comparison					
	0 vs. 1	0 vs. 2	0 vs. 3	1 vs. 2	1 vs. 3	2 vs. 3
No. of stools/day	.004	<.001	<.001	<.001	<.001	<.001
WBC	.16	<.001	<.001	.048	<.001	<.001
HGB	.11	<.001	<.001	.11	.001	.02
RBC	.3	.01	<.001	.24	<.001	.004
HCT	0.23	.002	<.001	.23	.001	.009
PLT	.34	<.001	<.001	.002	<.001	<.001
MONO	.45	<.001	<.001	.01	<.001	<.001
NEUT	.14	<.001	<.001	.047	<.001	.005
PCT	.35	<.001	<.001	.02	<.001	<.001
PDW	.7	.003	<.001	<.001	<.001	.45
PLCR	.33	<.001	<.001	.005	<.001	.33
MPV	.36	<.001	<.001	.003	<.001	.22
ESR	.25	<.001	<.001	.02	<.001	<.001
Fibrinogen mg/dl	.13	<.001	<.001	.006	<.001	<.001
CRP	.9	.047	<.001	.06	<.001	<.001
SI	.38	<.001	<.001	<.001	<.001	<.001
TP	.8	.005	<.001	.01	<.001	<.001
A1G	.1	<.001	<.001	.004	<.001	<.001

heatmap correlation matrix for all 18 numeric variables selected by the ANOVA method. Five strong correlations with a Pearson coefficient $\geq$ of 0.9 were identified between WBC and NEUT, PLT and PCT, HGB and HCT, PDW and MPV, PDW and PLCR. Thus, the following five parameters were removed from the analysis: NEUT, PCT, HCT, MPV and PLCR.

Chi-square test of independence identified seven categorical variables that are associated with Mayo score ($p \leq 0.01$): LGB, diarrhea, tenesmus, abdominal pain, weight loss, asthenia and pallor. Chi-square test results are given in Table 3.

As a result of the feature selection stage, 13 continuous parameters (WBC, PLT, MONO, ESR, fibrinogen, SI, TP, A1G, HGB, RBC, PDW, CRP, number of stools/day) and seven categorical parameters were included in further analysis.

3.4 Neural network models: Results

The initial dataset of 356 patient records was randomly divided into a training set (285 records) and test set (71 records) to build the classifiers. Thirty patient records were added independently to constitute the validation set. Unlike in the training and test sets, Mayo classes had a balanced distribution in the validation set. Patients' characteristics within the training, test and validation sets are illustrated in Table 4.

As a result of the feature selection step, three NN models were trained using the selected parameters as inputs. All three models were evaluated against several activation functions in order to maximize performance. The logistic function was selected as the transfer function of the hidden and output units due to highest performance. Supplementary Table 1 illustrates the best four activation functions (Logistic, TanH, Sinus, Elliott) and the corresponding accuracies obtained on each of the three classifiers. Other hyperparameters (such as the learning function and the parameters for the learning function) were selected so as to maximize performance. Table 5 summarizes all hyperparameters tuned for the construction of the classifiers.

The first NN model was developed using all 20 variables to predict whether a patient has endoscopic remission (Mayo 0 or 1) or active endoscopic disease (Mayo 2 or 3). The automatic tuning of the first classifier's hyperparameters resulted in one hidden layer containing one neuron. The model's performance metrics are shown in Table 6. On the train set, the model had an ACC of 92.63% (95% CI, 0.89–0.95; $p < .001$) with an SE of 90.22%, and SP of 93.78%, a PPV of 87.37% and NPV of 95.26% and an AUC of 0.92. On the test set, the model achieved a good performance with an ACC of 94.37% (95% CI, 0.862–0.9844; $p < .001$), an SE of 88%, an SP of 97.83%, a PPV of 95.65%, an NPV of 93.75% and an AUC of 0.9291. On the validation set, model predictions were similar in performance with ACC 93.33% (95% CI, 0.7793–0.9918;

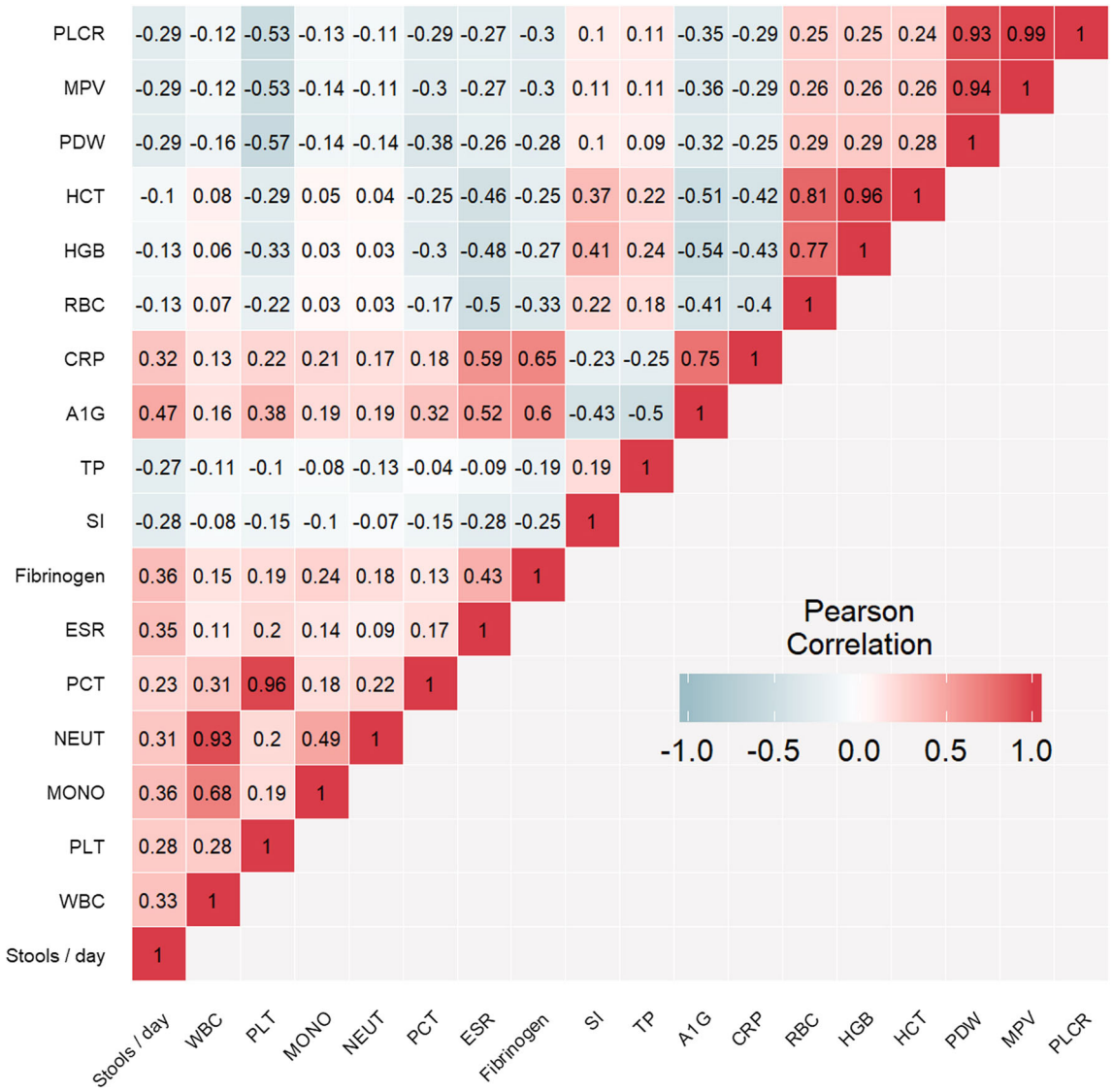

Figure 2 Legend: correlation heatmap showing the Pearson coefficients between all continuous parameters nominated by the feature selection method.

Table 3 Chi-square test for determining the association between categorical parameters and Mayo subscore (significance level: .01).

	Result	P value
Diarrhea	X^2 (3, N = 386) = 207.21	< .001
Tenesmus	X^2 (3, N = 383) = 127.97	< .001
LGB	X^2 (3, N = 385) = 223.08	< .001
Abdominal pain	X^2 (3, N = 384) = 99.578	< .001
Weight loss	X^2 (3, N = 382) = 64.43	< .001
Asthenia	X^2 (3, N = 384) = 89.186	< .001
Pallor	X^2 (3, N = 382) = 33.528	< .001
Smoking status	X^2 (6, N = 335) = 14.622	.02

$p < .001$), SE 92.86%, SP 93.75%, PPV 92.86%, NPV 93.75% and AUC 0.933. ROC curves proving model performance on the train, test and validation sets are shown in Figure 3.

The second NN model was built to predict the same binary outcome (endoscopic remission or activity) as the first classifier using only the 12 biological input parameters in order to investigate a model that uses only objective data. The hyperparameters of the second classifier were tuned automatically resulting in one hidden layer with three neurons. The second model's performance metrics are indicated in Table 7. On the train and test set, the model achieved a good performance: ACC 87% (95% CI, 0.8255–0.9069; $p < .001$), SE 89.13%, SP 86.01%, PPV 75.23%, NPV 94.32%, AUC 0.9084 on the train set and

Table 4 Clinical and biological parameters for train set, test set and validation set.

	Train set	Test set	Validation set
Number of patients	285	71	30
Gender (male:female)	195:90	46:25	16:14
Age (years)	45.7 ± 14	42.5 ± 13.5	41.1 ± 12.4
Number of stools/day	4.9 ± 3.7	5.4 ± 4.3	4.7 ± 4.5
LGB	194 (55%)	47 (66%)	16 (53%)
Diarrhea	197 (69.1%)	48 (67.6%)	20 (66.7%)
Tenesmus	137 (48.1%)	36 (50.7%)	11 (36.7%)
Abdominal pain	169 (59.3%)	45 (63.4%)	12 (40%)
Weight loss	90 (31.6%)	24 (33.8%)	4 (13.3%)
Asthenia	164 (57.5%)	40 (56.3%)	6 (20%)
Pallor	74 (26%)	19 (26.8%)	4 (13.3%)
WBC $*10^3/\mu L$	8.1 ± 2.8	8.5 ± 4.5	8.7 ± 3.6
HGB g/dL	13.3 ± 2.1	13.3 ± 1.9	13.6 ± 2
RBC $*100^3/\mu L$	4.7 ± 0.5	4.7 ± 0.5	4.8 ± 0.5
PLT $*10^3/\mu L$	308 ± 101.1	321.4 ± 132.5	308 ± 101.7
MONO $*10^3/\mu L$	0.67 ± 0.3	0.73 ± 0.4	0.7 ± 0.3
PDW fl	12.4 ± 2.2	12.3 ± 2.1	12.2 ± 2
ESR mm/1h	16.2 ± 19.7	17.4 ± 20.9	5.8 ± 4.2
Fibrinogen mg/dl	391.7 ± 78.3	395.2 ± 87.9	343.3 ± 98.9
CRP mg/dl	1.6 ± 3.2	2.1 ± 4.3	1 ± 1.7
SI μg/dl	67 ± 42.5	66 ± 34.7	73.4 ± 37.7
TP g/dl	7.4 ± 0.7	7.4 ± 0.7	7.4 ± 2.7
A1G %	2.7 ±1.1	2.9 ± 1.5	2.8 ± 1.3

Table 5 Tuning of the hyperparameters used for the construction of the neural network classifiers.

Parameters	st classifier	2nd classifier	3rd classifier
Number of input layer units	20	12	20
Number of hidden layers	1	1	1
Number of hidden layer units	1	3	5
Initialization function	Randomize weights		
Learning function	Standard Backpropagation		
Parameters for the learning function	$\eta = 0.2$ $d_{max} = 0$		
Update function	Topological Order		
Activation function for hidden and output units	Logistic		

η—the step width of the gradient descent; d_{max}—the maximum difference $d_j = t_j$-o_j between a teaching value t_j and an output o_j of an output unit which is tolerated.

ACC 88.73% (95% CI, 0.79–0.9501; $p < .001$), SE 88%, SP 89.13%, PPV 81.48%, NPV 93.18%, AUC 0.9191 on the test set. Performance on validation set was rather moderate: ACC 83.33% (95% CI, 0.6528–0.9436; $p < .001$), SE 78.57%, SP 87.5%, PPV 84.62%, NPV 82.35%, AUC 0.8482. Performance of the second classifier on the train, test and validation sets is shown in Figure 4.

The multiclass predictor (third NN model) was developed using all 20 variables to predict the endoscopic Mayo score. The model's output is a categorical variable with four possible values ranging from 0 to 3. The tuning of the third classifier's hyperparameters resulted in one hidden layer with five neurons. Model's renderings on each of the three datasets are shown in Table 8. The model had a good performance on the train set with an ACC of 89.44% 8304 (95% CI, 0.8526–0.9276; $p < .001$). and a multivariate predictor AUC of 0.9541. Model performance was moderate on the test set (ACC 76.06%, CI = 0.6446–0.8539; multivariate AUC 0.8726) and on the validation set (ACC 80%, CI = 0.6143–0.9229; multivariate AUC 0.9175). As defined by Hand and Till [20] as a mean of several AUC from pairwise comparison of classes, multiclass AUC was computed using the function multiclass.roc (pROC

Table 6 First classifier performance metrics.

Actual	Train set		Test set		Validation set	
	Predictions		Predictions		Predictions	
	Remission	Activity	Remission	Activity	Remission	Activity
Remission	83	12	22	1	13	1
Activity	9	181	3	45	1	15
ACC	92.63%		94.37%		93.33%	
95% CI	(0.8896, 0.9538)		(0.862, 0.9844)		(0.7793, 0.9918)	
P value	< .001		< .001		< .001	
SE	90.22%		88%		92.86%	
SP	93.78%		97.83%		93.75%	
PPV	87.37%		95.65%		92.86%	
NPV	95.26%		93.75%		93.75%	
AUC	0.92		0.9291		0.933	

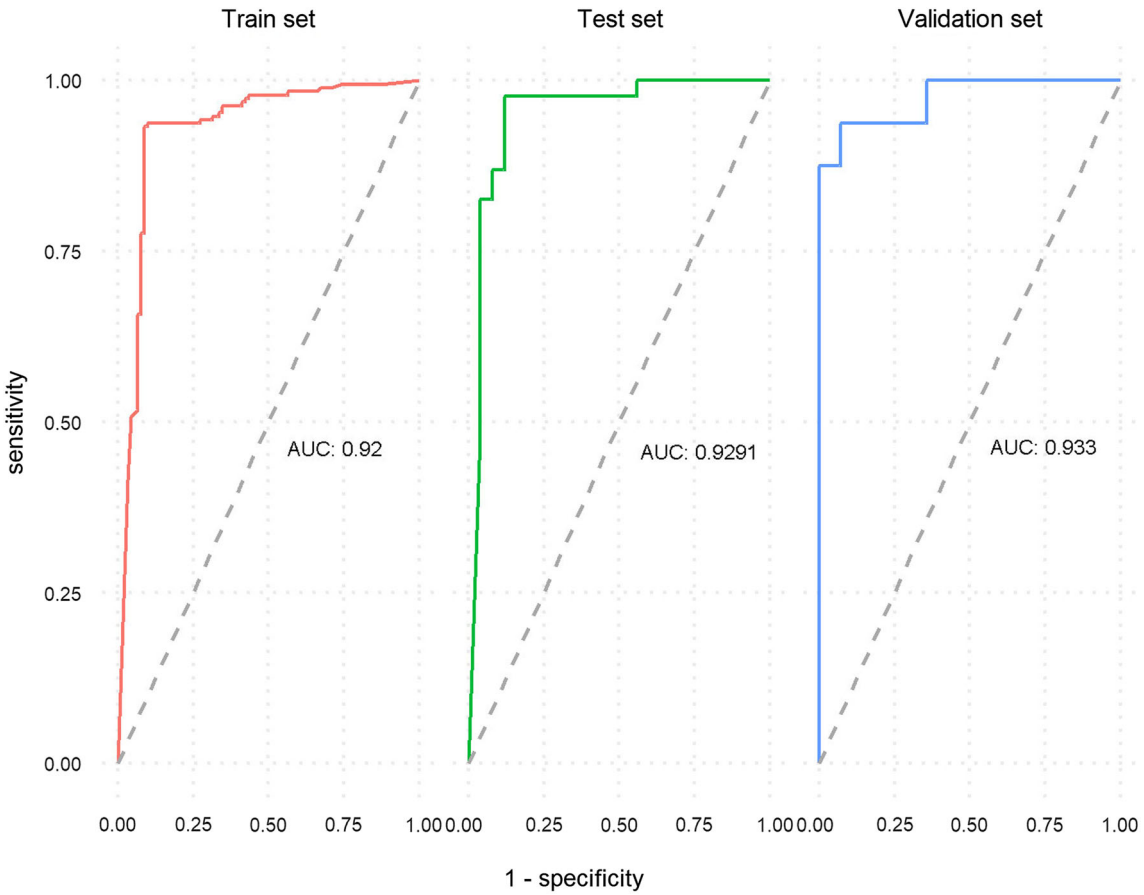

Figure 3 Legend: performance of the first classifier to predict endoscopic remission vs. relapse.

package in R). Multiclass AUC was calculated using a multivariate predictor with class probabilities.

The comparisons between the mlpML method and the rf, svmLinear and svmRadial are shown in Table 9. The rf, svmLinear and svmRadial methods were compared with each of the three proposed mlpML classifiers on the train, test and validation sets. Each algorithm is assessed based on the training time and accuracy of classification.

Table 7 Second classifier performance metrics.

Actual	Train set		Test set		Validation set	
	Predictions		Predictions		Predictions	
	Remission	Activity	Remission	Activity	Remission	Activity
Remission	82	27	22	5	11	2
Activity	10	166	3	41	3	14
ACC	87%		88.73%		83.33%	
95% CI	(0.8255, 0.9069)		(0.79, 0.9501)		(0.6528, 0.9436)	
P value	< .001		< .001		< .001	
SE	89.13%		88%		78.57%	
SP	86.01%		89.13%		87.5%	
PPV	75.23%		81.48%		84.62%	
NPV	94.32%		93.18%		82.35%	
AUC	0.9084		0.9191		0.8482	

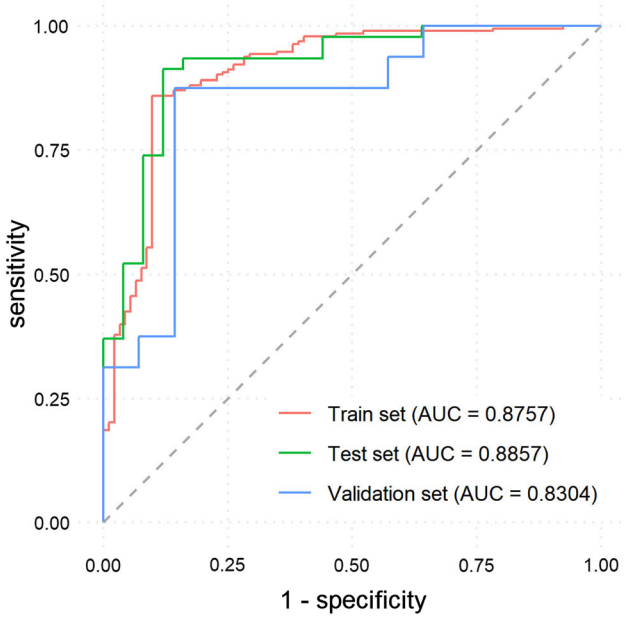

Figure 4 Legend: performance of the second classifier to predict endoscopic remission vs. relapse.

4 Discussions

Our study is the first neural network developed for predicting endoscopic disease activity in UC based on routinely available clinical parameters. We demonstrated that using only standard, non-costly, non-invasive clinical and biological data can differentiate active UC from inactive UC.

To date, numerous studies have been conducted in order to establish non-invasive biomarkers capable of estimating endoscopic activity in UC, avoiding invasive diagnostic tests [21]. Of all, fecal markers (especially calprotectin) were noted to yield the best performance so far, with high accuracy and sensitivity [22]. Thereby, fecal calprotectin (FC) is now used in UC patients' clinical management according to several consensus guidelines [5, 23]. However, FC has not been validated for biomarker-based decision making due to low reliability: significant variability across platforms (which makes it difficult to establish a cutoff value), intra-individual day-to-day variation FC concentrations, degradation of FC levels at room

Table 8 Multiclass predictor performance (third classifier).

	Train set				Test set				Validation set			
	Mayo predictions				Mayo predictions				Mayo predictions			
Actual Mayo	0	1	2	3	0	1	2	3	0	1	2	3
0	69	2	1	0	12	1	0	0	8	1	1	0
1	2	65	8	0	3	7	2	1	0	5	1	0
2	0	3	58	9	0	1	23	2	0	0	5	1
3	0	1	4	62	0	1	6	12	0	0	2	6
ACC	89.44%				76.06%				80%			
95% CI	(0.8526, 0.9276)				(0.6446, 0.8539)				(0.6143, 0.9229)			
P value	< .001				< .001				< .001			
Multivariate predictor AUC	0.9541				0.8726				0.9175			

Table 9 Comparisons between the mlpML (multilayered perceptron) and other methods of the caret::train function: rf (random forest), svmLinear (support vector machine with linear kernel), svmRadial (support vector machine with radial basis function kernel) concerning the training time, hyperparameters and accuracy

	Method	mlpML	rf	svmLinear	svmRadial	
1st classifier	Training time (sec)	62.49	134.06	7.89	24.83	
	Tuned hyperparameters	1 hidden unit (1 layer)	mtry = 2	C = 1	sigma = 0.0433973 C = 0.25	
	ACC	Train set	92.63%	99%	88.4%	90%
		Test set	94.37%	94%	86%	92.9%
		Validation set	93.33%	90%	83%	90%
2nd classifier	Training time (sec)	54.1	120.58	6.25	24.5	
	Tuned hyperparameters	3 hidden units (1 layer)	mtry = 2	C = 1	sigma = 0.1190891 C = 0.5	
	ACC	Train set	87%	99%	80%	82%
		Test set	88.73%	76%	80%	70.4%
		Validation set	83.33%	83%	73%	76.7%
3rd classifier	Training time (sec)	32.19	73.3	3.23	11.95	
	Tuned hyperparameters	5 hidden units (1 layer)	mtry = 2	C = 1	sigma = 0.04594297 C = 1	
	ACC	Train set	89.44%	100%	80%	84.5%
		Test set	76.06%	69%	69%	64%
		Validation set	80%	73%	63%	70%

temperature after stool collection [22]. Additionally, FC is not allowed in Asian and some Western countries [24] and is more expensive than routine laboratory investigations. On the contrary, classical and emerging serum biomarkers, although more stable and with less intra- and inter-variability than fecal markers, proved lower accuracy when examined separately [25].

Individual predictive powers of classical markers, although small, were considered together in a composite approach through our ML solution, yielding significantly better results. Thus, the feature selection step consisting of ANOVA pairwise comparisons resulted in 13 continuous parameters (WBC, PLT, MONO, ESR, fibrinogen, SI, TP, A1G, HGB, RBC, PDW, CRP, number of stools/day) and seven categorical parameters to be used as inputs for our ML solution. Our selected biological parameters' clinical relevance is confirmed by numerous studies in which standard inflammatory biomarkers such as WBC, CRP, ESR, fibrinogen, PLT and A1G have already been studied as markers for disease activity in UC [26]. Moreover, abnormal values of HGB, RBC, SI are a consequence of gastrointestinal bleeding that is one of the main symptoms of active UC. HGB has also been found to correlate with endoscopic activity [27]. The clinical parameters, such as the number of stools/day and all categorical parameters selected as inputs for the NN algorithm, reflect various manifestations of active UC. The rationale for including clinical data in our model is that several studies assessing composite markers consisting of clinical and biological data have shown superiority over biological data alone in discriminating active and inactive IBD [28].

The first classifier proposed in our study used eight clinical parameters and 12 routine biological tests to differentiate endoscopic active UC from inactive UC with a high ACC of 94.37%, and SE of 88%, an SP of 97.83% and an AUC of 0.9291 on the test set and ACC 93.33%, SE 92.86%, SP 93.75% and AUC 0.933 on the validation set.

A recent meta-analysis that evaluated FC in assessing IBD endoscopic activity reported a pooled sensitivity of 87.3% (85.4–89.1), specificity of 77.1% (73.7–80.3) and AUC of 0.91 in UC [29]. Our study's ML approach achieved higher performance than all indicators reported in the meta-analysis.

It is essential to notice that studies evaluating non-invasive methods to assess UC endoscopic activity repeatedly use the binary classification (Mayo 0 and 1 for inactive/Mayo 2 and 3 for active disease) rather than the complete endoscopic Mayo score. This approach is justified since a favorable prognosis was documented for patients with a Mayo score of 0 and 1 (more prolonged clinical remissions and lower rates of colectomy) than the patients with a Mayo score of 2 and 3 [30]. This enhances the idea that an ML model like the one advanced in our study, with an excellent performance on the test and validation sets, will prove to be a useful tool for disease monitoring in clinical practice after rigorous validations on external patient datasets.

The second classifier presented in our paper used only the 12 routine biological tests to predict the same binary

output that differentiates inactive from active disease with an ACC 88.73%, SE 88%, SP 89.13%, AUC 0.8857 on the test set, and ACC 83.33%, SE 78.57%, SP 87.5%, AUC 0.8304 on the validation set. The results obtained on the test set are similar to the results of the meta-analysis mentioned above (a SE of 88% in our study vs. a pooled SE of 87.3% for FC in the meta-analysis). However, our classifier obtained higher values for SP (89.13% in our study vs. 77.1% for FC in the meta-analysis). Additionally, a high NPV of 93.18% obtained by the second NN model on the test set could reliably rule out active UC. The performance metrics obtained on the validation set were slightly lower, except for SP, which was higher than the pooled results of FC reported in the meta-analysis. Thereby, the second model could successfully be used in clinical practice, considering that the results of the second classifier (ACC 88.73% on the test set) are roughly similar to the FC performance in detecting active disease along with the fact that the use of FC is now recommended in the management of UC patients in several consensus guidelines: "FC levels correlate with degrees of endoscopic and histologic inflammation in UC and therefore have been proposed as a marker of disease activity to guide treatment. FC levels are more sensitive and specific than serum inflammatory markers and also less invasive than endoscopy or mucosal biopsies, so this assessment has become routine for many clinicians who are managing patients with UC" [5, 23]. Moreover, considering that our model uses only cheap, more stable, immediately available laboratory tests than FC, future research on larger datasets remains of interest for further validation of the proposed classifier.

Third classifier used all 20 parameters to predict the endoscopic Mayo score with an ACC of 76.06% (95% CI, 0.6446–0.8539; $p < .001$) and a multivariate AUC of 0.8726 on the test set and an ACC of 80% (95% CI, 0.6143–0.9229; $p < .001$) and multivariate AUC of 0.9175. To date, there are no studies that aimed to predict the endoscopic Mayo score using only non-invasive biomarkers. Although binary classification into active/inactive disease is of compelling importance, there are studies to suggest that differentiating between endoscopic Mayo 0 and 1 may have a better impact on predicting future outcomes. In a study, the risk of relapse for UC patients was higher for a Mayo score of 1 than Mayo 0 [31]. For this reason, results obtained by the multiclass predictor in our study could open the path for further research in estimating complete endoscopic Mayo subscore by non-invasive ML methods, taking into account that although ACCs obtained were moderate (but not small), the higher multiclass AUCs keep the hopes high.

Three of the most common algorithms used in healthcare applications are the neural networks, support vector machines and random forests [32]. In terms of comparisons between the different algorithms, the results illustrated in Table 9 justify our choice for the multilayered perceptron method. The random forest approach was both slower and less accurate, with a tendency to overfit. The support vector machine methods were undoubtedly faster at the cost of significantly lower accuracies.

4.1 Limitations and future perspectives

Firstly, our dataset's small size and the fact that the independent validation set is from a single center entail rigorous external validation with data from other centers. Secondly, the imbalanced distribution of Mayo classes in the initial dataset of 356 records (with Mayo 2 class containing three times more records than each of the other groups) predisposes to calculation biases. However, the SMOTE function in R was used to reduce these biases significantly. Thirdly, FC was not documented for a direct real-time comparison with the proposed NN classifiers' performance.

In the future, these drawbacks could be overcome by employing studies on a larger number of patients in a center with greater accessibility that would permit organizing a cohort with a balanced distribution of Mayo classes. Next trials would improve models' performance by incorporating more ML algorithms and real-time comparisons with the documented FC levels.

Further improving and validating automatic learning methods in this area may lead to more frequent monitoring of UC patients, significantly fewer invasive procedures, less exposure to inherent risks and more comfort for the patients. Intensified UC monitoring may lead to the early tracing of subclinical inflammation. Early detection of inflammatory relapses is of great importance since long-lasting subclinical disease activity increases the risk of colonic neoplasia and decreases patients' quality of life and productivity [1, 33].

For our method to make its way into clinical practice and follow the path of other AI solutions approved by the Food and Drug Administration (Guardian Connect System [34], WAVE Clinical Platform [35]), future head to head prospective trials to compare the performance of AI models and fecal calprotectin, randomized trials between the two techniques and the approval of a national or international regulatory body will be statutory.

5 Conclusions

In the context of a difficult to manage disorder with unknown etiology and pathogenesis [36], unpredictable disease course and current invasive diagnostic methods, our study proposes a cost-efficient, non-invasive

technique to predict UC activity accurately using modern computational solutions in the area of AI/ML. Our neural network model represents a significant advance in the non-invasive assessment of inflammation in UC, leading to further research and possible future use in clinical practice. At a time when neural network algorithms are becoming increasingly sophisticated, less complex neural networks, such as the multilayered perceptron, are able to solve real-world medical problems in a time-efficient manner and based on reduced amounts of data.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [Iolanda V. Popa], [Otilia Gavrilescu] and [Mihaela Dranga]. The first draft of the manuscript was written by [Iolanda V. Popa], and all authors commented on previous versions of the manuscript. [Iolanda V. Popa], [Cristina Cijevschi Prelipcean] helped in conceptualization; [Iolanda V. Popa, Alexandru Burlacu] contributed to methodology; [Iolanda V. Popa] involved in formal analysis and investigation; writing—original draft preparation; [Alexandru Burlacu], [Iolanda V, Popa] helped in writing—review and editing; [Cristina Cijevschi Prelipcean], [Cătălina Mihai] helped in resources; [Alexandru Burlacu], [Cătălina Mihai] supervised. All authors read and approved the final manuscript.

Funding No funding declared.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were by the ethical standards of the local/regional Romanian institutions ("St. Spiridon" Regional Hospital Ethics Committee, no. 54/10.2019 and Research Ethics Commission of the "Gr. T. Popa" University of Medicine and Pharmacy, no. 15308/07.2019) and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Habibi F, Habibi ME, Gharavinia A, Mahdavi SB, Akbarpour MJ, Baghaei A, Emami MH (2017) Quality of life in inflammatory bowel disease patients: A cross-sectional study. Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences
2. Marrero F, Qadeer MA, Lashner BA (2008) Severe Complications of Inflammatory Bowel Disease. Med Clin North Am 92(3):671–686
3. Burisch J, Jess T, Martinato M, Lakatos PL (2013) The burden of inflammatory bowel disease in Europe. J Crohns Colitis 7(4):322–337
4. Annese V, Daperno M, Rutter MD, Amiot A, Bossuyt P, East J, Ferrante M, Götz M, Katsanos KH, Kießlich R, Ordás I, Repici A, Rosa B, Sebastian S, Kucharzik T, Eliakim R (2013) European evidence based consensus for endoscopy in inflammatory bowel disease. J Crohn's Colitis 7(12):982–1018
5. Lamb CA, Kennedy NA, Raine T, Hendy PA, Smith PJ, Limdi JK, Hayee BH, Lomer MC, Parkes GC, Selinger C, Barrett KJ (2019) British Society of Gastroenterology consensus guidelines on the management of inflammatory bowel disease in adults. Gut 68(3):1–06
6. Magro F, Gionchetti P, Eliakim R, Ardizzone S, Armuzzi A, Barreiro-de Acosta M, Burisch J, Gecse KB, Hart AL, Hindryckx P, Langner C (2017) Third European evidence-based consensus on diagnosis and management of ulcerative colitis. Part 1: definitions, diagnosis, extra-intestinal manifestations, pregnancy, cancer surveillance, surgery, and ileo-anal pouch disorders. J Crohn's Colitis 11(6):649–670
7. Ko CW (2018) Colonoscopy Risks: What Is Known and What Are the Next Steps? Gastroenterology 154(3):473–475
8. Cogan T, Cogan M, Tamil L (2019) MAPGI: Accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning. Comput Biol Med 111:103351
9. Maeda Y, Kudo S, Mori Y, Misawa M, Ogata N, Sasanuma S, Wakamura K, Oda M, Mori K, Ohtsuka K (2019) Fully automated diagnostic system with artificial intelligence using endocytoscopy to identify the presence of histologic inflammation associated with ulcerative colitis (with video). Gastrointest Endosc 89(2):408–415
10. Biasci D, Lee JC, Noor NM, Pombal DR, Hou M, Lewis N, Ahmad T, Hart A, Parkes M, McKinney EF, Lyons PA, Smith KG (2019) A blood-based prognostic biomarker in IBD. Gut 68(8):1386–1395
11. Waljee AK, Liu B, Sauder K, Zhu J, Govani SM, Stidham RW, Higgins PD (2018) Predicting corticosteroid-free endoscopic remission with vedolizumab in ulcerative colitis. Aliment Pharmacol Ther 47(6):763–772
12. Hardalaç F, Basaranoglu M, Yuksel M, Kutbay U, Kaplan M, Ozderin Ozin Y (2015) The rate of mucosal healing by azathioprine therapy and prediction by artificial systems. Turk J Gastroenterol 26(4):315–321
13. Popa IV, Burlacu A, Mihai C, Prelipcean CC (2020) A machine learning model accurately predicts ulcerative colitis activity at one year in patients treated with anti-tumour necrosis factor α agents. Medicina 56(11):628
14. Mossotto E, Ashton JJ, Coelho T, Beattie RM, MacArthur BD, Ennis S (2017) Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. Sci Rep 7(1):1–10
15. Do Q, Son TC, Chaudri J (2017) Classification of Asthma Severity and Medication Using TensorFlow and Multilevel Databases. Procedia Comput Sci 113:344–351
16. Sideris C, Shahbazi B, Pourhomayoun M, Alshurafa N, Sarrafzadeh M (2014) Using electronic health records to predict severity of condition for congestive heart failure patients. InProceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (pp. 1187-1192)
17. Le S, Hoffman J, Barton C, Fitzgerald JC, Allen A, Pellegrini E, Calvert J, and Das R (2019) "Pediatric severe sepsis prediction using machine learning." Frontiers in Pediatrics 7
18. Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Stat Surv 4:40–79
19. Sood R (2018) *Comparative Data Analytic Approach for Detection of Diabetes. Doctoral dissertation,* University of Cincinnati
20. Hand DJ, Till RJ (2001) A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn 45(2):171–186

21. Nakov R (2019) New markers in ulcerative colitis. Clin Chim Acta 497:141–146
22. Dulai PS, Peyrin-Biroulet L, Danese S, Sands BE, Dignass A, Turner D, Mantzaris G, Schölmerich J, Mary JY, Reinisch W, Sandborn WJ (2019) Approaches to integrating biomarkers into clinical trials and care pathways as targets for the treatment of inflammatory bowel diseases. Gastroenterology 157(4):1032-1043.e1
23. Rubin DT, Ananthakrishnan AN, Siegel CA, Sauer BG, Long MD (2019) ACG clinical guideline: Ulcerative colitis in adults. Am J Gastroenterol 114(3):384–413
24. Shiga H, Abe I, Onodera M, Moroi R, Kuroha M, Kanazawa Y, Kakuta Y, Endo K, Kinouchi Y, Masamune A (2020) Serum C-reactive protein and albumin are useful biomarkers for tight control management of Crohns disease in Japan. Sci Rep 10(1):1–8
25. Nielsen OH, Vainer B, Madsen SM, Seidelin JB, Heegaard NHH (2000) Established and emerging biological activity markers of inflammatory bowel disease. Am J Gastroenterol 95(2):359–367
26. Fengming Y, Jianbing W (2014) Biomarkers of inflammatory bowel disease. Disease Markers 2014
27. Miranda-García P, Chaparro M, Gisbert JP (2016) Correlation between serological biomarkers and endoscopic activity in patients with inflammatory bowel disease. Gastroenterol Hepatol 39(8):508–515
28. Dragoni G, Innocenti T, Galli A (2020) Biomarkers of inflammation in Inflammatory Bowel Disease: how long before abandoning single-marker approaches?. Digestive Diseases
29. Rokkas T, Portincasa P, Koutroubakis IE (2018) Fecal Calprotectin in Assessing Inflammatory Bowel Disease Endoscopic Activity: a Diagnostic Accuracy Meta-analysis. J Gastrointestin Liver Dis 27(3):299–306
30. Picco MF (2017) PiCaSSO: a predictive score for endoscopic findings in ulcerative colitis that sounds like art but is all science. Gastrointest Endosc 86(6):1128–1130
31. Barreiro-de Acosta M, Vallejo N, de la Iglesia D, Uribarri L, Bastón I, Ferreiro-Iglesias R, Lorenzo A, Domínguez-Muñoz JE (2016) Evaluation of the Risk of Relapse in Ulcerative Colitis According to the Degree of Mucosal Healing (Mayo 0 vs 1): A Longitudinal Cohort Study. J Crohn's Colitis 10(1):13–19
32. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y (2017) Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology 2(4):230–243
33. Cleveland NK, Rubin DT, Hart J, Weber CR, Meckel K, Tran AL, Aelvoet AS, Pan I, Gonsalves A, Gaetano JN, Williams K, Wroblewski K, Jabri B, Pekow J (2018) Patients with ulcerative colitis and primary sclerosing cholangitis frequently have subclinical inflammation in the proximal colon. Clin Gastroenterol Hepatol 16(1):68–74
34. Benjamens S, Dhunnoo P, Meskó B (2020) The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 3(1):1–8
35. Hamamoto R, Suvarna K, Yamada M, Kobayashi K, Shinkai N, Miyake M, Takahashi M, Jinnai S, Shimoyama R, Sakai A, Takasawa K (2020) Application of artificial intelligence technology in oncology: Towards the establishment of precision medicine. Cancers 12(12), 3532
36. Kumar M, Garand M, Al Khodor S (2019) Integrating omics for a better understanding of Inflammatory Bowel Disease: a step towards personalized medicine. J Transl Med 17(1):1–13