Classification and prediction of inflammatory bowel diseases

J.Hema sri , K.Hemanth
V.Hemasree, B.Hema Vardhan
J.Nagateja , B.Hima Vaishnavi

Guide : Shiva Kumar
Professor

Artifical Intelligence&Machine Learning
Department Of Computer Science & Engineering
Malla Reddy University
Hyderabad , Telangana ,India

Abstract: The project aims to develop a system For predicting inflammatory bowel and building the machine learning model based on routinely performed laboratory blood, urine, and fecal tests to support differentiation between IBD patients and non-IBD patients comparison of the effectiveness of our model to standard inflammatory serum marker, that is C-reactive protein (CRP), in the prediction of IBD, creating a website- based application supporting the prediction of the presence of IBD . the age profile of IBD patients is changing and there is an increase in early-onset and late-onset IBD prevalence. Both groups (older adults, which frequently suffer from various comorbidities, as well as children) would particularly benefit from the non-invasive diagnostic test. However, colonoscopy remains a gold standard in IBD diagnosis, monitoring of the disease course, and response to the therapy, as well as colorectal cancer screening [6,7,8,9]. Still, despite its obvious advantages, it is highly invasive, expensive, time-consuming, requires qualified medical personnel and patient‘s preparation, and is often poorly tolerated by patients themselves. Besides, in the pandemic all low-contact medical procedures are preferred. creating a website- based application supporting the prediction of the presence of IBD.Therefore, a simple diagnostic methodology based only on markers from blood, urine and stool that can be performed by a GP would be imperative inthe early diagnosis of IBD

Keywords :
inflammatoryboweldisease;ulcerative colitis; Crohn’s disease; artificial intelligence; machine learning; model; prediction

I. INTRODUCTION

Inflammatory bowel disease (IBD) is a chronic, incurable disease of the gastrointestinal tract represented by two most common forms: ulcerative colitis (UC) and Crohn’s disease (CD). The pathogenesis of IBD is not fully explained, and according to the “IBD interactome” concept, it involves interrelation between genetic, microbiological, environmental, and immune factors [1,2]. Though unclear pathogenesis results in the lack of effective therapeutic modalities and the absence of efficient prevention, this in the light of the increasing prevalence of IBD worldwide has a particular meaning. Consecutively, there is no single test sufficient to provide a diagnosis. Recognition of IBD is based on a combination of clinical symptoms, laboratory tests, and endoscopic and imaging tests together with pathological examination [3,4]. Remarkably, despite progress in endoscopic and imaging techniques, the Bioethics Committee of the Wroclaw Medical University N° KB-504/2021objectives, we took an attempt to analyse with machine learning models simple laboratory tests performed in real clinical life. Most robust classificators belonging to the random forest family obtained 97% and 91% mean average precision for Crohn’s disease and ulcerative colitis, respectively. The feasibility of making UC diagnoses using non-invasive methods is possible by the random forest classifier we selected, which achieved satisfactory results, when matching to age, gender, and 14 laboratory features that can be easily verified by a general practitioner: P-

DOI: 10.55041/IJSREM27527

LCR, ESR, fecal calprotectin, haemoglobin, creatinine, MCH, peripheral blood leukocytes, cholesterol-LDL, peripheral blood erythrocytes, CEA, bacteria in urine, glucose in urine, microscopic stool ova and parasites test, and HBeAg. In turn for CD, the random forest algorithm showed that the most significant attributes were age, gender, and further laboratory markers: MCH, MPV, MCHC, peripheral blood neutrophils, total bilirubin, HCT, potassium, AST, AP, peripheral blood monocytes, erythrocytes, basophils, and erythroblasts.It has to be emphasized that both, developed model and web application, require further research in the large cohort of patients with suspicion of IBD and comparison to endoscopy as a reference test.Multiple biomarkers, which have been identified in presented models as important disease descriptors, have pathogenetic connection with the inflammatory bowel disease, mainly in respect to indices of inflammation, anemia, andmalnutrition.

II. PROBLEM STATEMENT

The goal of treatment is to induce remission for either UC or CD. Treatment of IBD is divided into the management of mild, moderate, and severe disease. Agents formerly reserved for the more severe disease are now employed sooner. UC treatment depends greatly on the extent of the disease and the presence of extraintestinal manifestations. For those with mild to moderate disease limited to the rectum, aminosalicylate agents like mesalamine are the mainstays. Mesalamine is administered rectally but may be combined with oral therapy to induce or maintain remission. For those patients with moderate disease who are refractory to mesalamine, oral glucocorticoids or immunomodulators such as TNF-alpha monoclonal antibodies (infliximab) may be an option. Up to 25% of all UC patients will require total colectomy for the uncontrolled disease. Proctocolectomy with ileal pouch-anal anastomosis (IPAA) is the procedure of choice for elective cases.[9][10][11]

1.16 significant morbidity in these patients. If steroid use for more than three months is expected, then calcium supplements and bisphosphonatesshould be introduced.

III. METHODOLOGY

The input data was initially divided into learning and test sets with a 7:3 ratio due to small sample size. Then, scaling and dimensionality reduction of the data were performed using the principal component analysis (PCA) method. About 443 primary extracted features were reduced to 64 unique instances since the datasets differed in incomplete data threshold. Most of the samples in all groups (UC-study, CD-study, or control) were not complete, so the obtained data were also subjected to data supplementation by multiple imputation method using IterativeImputer from the Scikit-learn library [12]. Specifically, for each feature (physical or laboratory marker, as shown in Table 3), the median was calculated on the observed data, which was decreased or increased by a random valuefrom the interval (where SD is the standard deviation calculated on the observed samples) and then iteratively entered in each cell with the missing value. Such an operation was performed separately for each feature and separately for the studied and control groups. Exceptions are values that have been converted from text data. For such data, an integer is drawn from the given range with appropriate weights, which were calculated by dividing the number of occurrences of unique values by the total number of feature values. In the first step, machine learning (ML) classifiers with high prediction were run through hyperparameters tuning. These classifiers included logistic regression, k-nearest neighbor, decision tree forests, support vector machine, and gradient boosting, with different hyperparameters. Next, a majority voting classifier was introduced for each set, based on ensemble learning, which combines several models to create a single most optimal predictive model. Automatic tuning of hyperparameters was performed using the grid method with greedy algorithm.

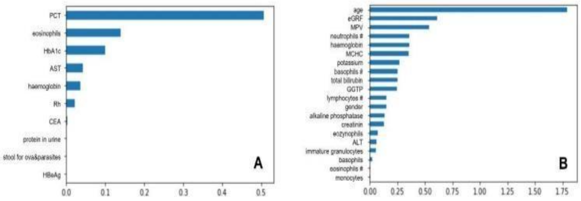
The algorithm greedily selects thefeatures that give it the best result. It checked between 20 and
10 featurings. Each classifier can thus have optimally chosen the number of features that gives the best result. The whole process was accompanied by cross-validation with a k parameter equal to 10, which was proven to give the best variance to load trade-off. Evaluating the Effectiveness of the Model In order to correctly select classifiers, a mechanism for estimating performance had to be provided. Without introducing a way to evaluate the model, it would not be possible to determine the optimal balance between variance and load. Every model is exposed to the risk of under-fitting (i.e., high loading) due to low complexity or over-training (i.e., high variance due to too high co

EXPERIMENTAL RESULTS:

3.1. Data Filtering and Input Features The processing operation of the provided data resulted in 443 data features, which represented the unique biological or blood-, urine-, or stool-based laboratory parameter of the patient. Unfortunately, most of these laboratory tests were performed on few patients only and do not represent a reliable source of information by their presence at less than 30% in a given group (UC-study, CD-study, or control). Filtering these data ultimately yielded 64 features present in all groups (Table 3), from which the data analysis was then performed. It should also be noted that in addition to the total number of white blood cells, morphological studies also report the expanded blood pattern, that is, the number of each type of white blood cell per unit volume. The duplicate feature names are not an error but a distinction, since for each total blood cell count (indicated in Table 3 with the ‘#’ sign), a percentage (‘%’ sign in Table 3) within theexpanded blood pattern has also been calculated.

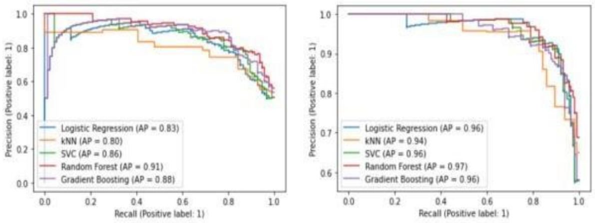
3.2. Machine Learning Classifiers Each machine learning task is to work out a certain solution with the help of an appropriate mathematical model, whose parameters are not known at the beginning. At the input of the model, we have data which are the specific values of certain features, while at the output of the model we get the solution to the task associated with certain domain objects as individuals, examples, specimens, measurements in the world. Here, the features are laboratory markers obtained from historical patients. The problem of machine learning is the automatic, machine building of the model with the help of an appropriate algorithm. Here the following classification algorithms were used.

3.2.1. Logistic Regression For the set of classifiers, multiple variants of logistic regression were tested for optimization and to find the best combinations of parameters such as optimization algorithm, fitting method, weights, regularization penalties, and the so- called C parameter. The grid method showed that the worst prediction results were achieved by classifiers with the regularization set to L1 (internal algorithm setting under Python programing language) and the inverse of the regularization strength equal to
0.01. No significant changes were observed between the default value of C and 0.1, and finally C value of 0.1 was considered the optimal value. Different values of the optimization algorithm, by themselves, did not affect prediction performance, suggesting their importance only when combined with other parameters. Models with a linear optimization algorithm and both variants of the fitting method received good results. Finally, the following hyperparameters were found to be the optimal for UC dataset:
{‘classifier__C’: 0.01,
‘classifier__penalty’: ‘l2’,
‘classifier__solver’: ‘newton-cg’,
‘featureSelector_
_n_features_to_select’: 10}, and the most relevant were following features: plateletcrit (PCT), eosinophils, haemoglobin A1c (HbA1c), aspartate aminotransferase (AST), haemoglobin, Rhesus (Rh), carcinoembryonic antigen (CEA), protein in urine, microscopic stool for ova and parasites test, and Hepatitis B e Antigen (HBeAg). Meanwhile, for CD dataset the most optimal clasificator parameters were
{‘classifier__C’: 0.1,
‘classifier__penalty’: ‘l2’,
‘classifier__solver’: ‘newton-cg’,
‘featureSelector_
_n_features_to_select’: 20}



3.3. Best Classifiers and Most Important Predictors

The initial prediction was made for 64 features, shown in Table 3 which means that a high prediction score can be achieved for a patient with 62 tests performed and age and gender completed. This is not an economical solution from the GP's point of view. Therefore, cut-off thresholds between 10 and 20 features of significance were introduced. This means that the prediction accuracy was checked for 20, 19, 18, . . . , 10 features, respectively. The prediction improved for the 15 or 16 cut-off thresholds, but decreased again for the lower and higher number of important features. This may indicate a negative effect of insignificant features on prediction quality, but also a negative effect of too few features. The more insignificant features we remove, the better the prediction for the disease entity and the smaller the standard deviation. However, the removal should not be exaggerated, as too large a cutoff may lead to a deterioration in model quality. Finally, for the UC case, 16 features resulted as an optimal value, while for the CD case 15 features the best optimized prediction quality of the model. Both models were based on random forests classifier.



(a) (b)

CONCLUSION:

Complex and unclear pathogenesis of inflammatory bowel disease stays behind the failure into identifying a single biomarker of disease. As IBD has become a global disease, simple and available diagnostic tools are essential. Therefore, we attempted to create a machine-learning algorithm to support IBD diagnosis. Results of our pilot study suggest that routine blood, urine and fecal markers based machine learning model may supportwith high accuracy, higher than CRP, the diagnosis of IBD. However, validation of the test is vital before it can be considered to be applied in clinical practice. we took an attempt to analyse with machine learning models simple laboratory tests performed in real clinical life. Most robust classificators belonging to the random forest family obtained 97% and 91% mean average precision for Crohn's disease and ulcerative colitis, respectively. The feasibility of making UC diagnoses using non-invasive methods is possible by the random forest classifier we selected, which achieved satisfactory results, when matching to age, gender, and 14 laboratory features that can be easily verified by a general practitioner: P-LCR, ESR, fecal calprotectin, haemoglobin, creatinine, MCH, peripheral blood leukocytes, cholesterol-LDL, peripheral blood erythrocytes, CEA, bacteria in urine, glucose in urine, microscopic stool ova and parasites test, and HBeAg. In turn for CD, the random forest algorithm showed that the most significant attributes were age, gender, and further laboratory markers: MCH, MPV, MCHC, peripheral blood neutrophils, total bilirubin, HCT, potassium, AST, AP, peripheral blood monocytes, erythrocytes, basophils, and erythroblasts. It has to be emphasized that both, developed model and web application, require further research in the large cohort of patients with suspicion of IBD and comparison to endoscopy as a reference test. Multiple biomarkers, which have been identified in presented models as important disease descriptors, have pathogenetic connection with the inflammatory bowel disease, mainly in respect to indices of inflammation, anemia, and malnutrition. First of all, ESR and CRP represent classic serum inflammatory markers [22], whereas fecal calprotectin reflects gastrointestinal inflammation [23]. CRP is an acute-phase reactant produced by hepatocytes in response to stimulation from inflammatory cytokines such as interleukin- 1, interleukin-1β, and tumor necrosis factor- alpha [24], whereas, ESR indicates the migration speed of red blood cells in plasma

[25] Meta-analysis aimed to assess the utility of CRP, ESR, FC, and fecal lactoferrin to exclude inflammatory bowel disease in adults with IBS demonstrated that at a CRP level of ≤0.5 or calprotectin level of ≤40 µg/g, there was a ≤1% probability of having IBD. However, individual analysis of ESR had little clinical utility [26]. It has to be emphasized that measurement of fecal calprotectin is a noninvasive test that has an established position in the differentiation between inflammatory and non- inflammatory gastrointestinal conditions, for instance, irritable bowel syndrome (IBS) as well as monitoring of the IBD course [27]. Yet, the assessment of fecal calprotectin concentration has also several limitations. For instance, it may be influenced by:
gastrointestinal infections, gastric and colonic malignancies, eosinophilic colitis, lymphocytic colitis, and coeliac disease,

- concomitant medical treatment with proton pump inhibitors, nonsteroidal anti-inflammatory drugs, and acetylsalicylic acid,
- age.

REFERENCES:
1. De Souza, H.S.P.; Fiocchi, C.; Iliopoulos, D. The IBD interactome: An integrated view of etiology, pathogenesis, and therapy. Nat. Rev. Gastroenterol. Hepatol. 2017, 14, 739–749. [CrossRef] [PubMed] 2. Kellermayer, R.; Zilbauer, M. The gut microbiome and the triple environmental hit concept of, A.; Barreiro-de Acosta, M.; Burisch, J.; Gecse, K.B.; Hart, A.L.; Hindryckx, P.; et al. 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. Crohns Colitis 2017, 11, 649–670. [CrossRef] 4. Gomollón, F.; Dignass, A.; Annese, V.; Tilg, H.; Van Assche, G.; Lindsay, J.O.; Peyrin-Biroulet, L.; Cullen, G.J.; Daperno, M.; Kucharzik, T.; et al. 3rd European evidence-based consensus on the diagnosis and management of Crohn's disease 2016: Part 1: Diagnosis and medical management. J. Crohns Colitis 2017, 11, 3–25. [CrossRef] 5. Cantoro, L.; Di Sabatino, A.; Papi, C.; Margagnoni, G.; Ardizzone, S.; Giuffrida, P.; Giannarelli, D.; Massari, A.; Monterubbianesi, R.; Lenti, M.V.; et al. The time course of diagnostic delay in inflammatory bowel disease over the last 436 sixty years: An Italian multicentre study. J. Crohns Colitis 2017, 11, 975–980. [CrossRef] 6. Dave, M.; Loftus, E.V., Jr. Mucosal healing in inflammatory bowel disease-a true paradigm of success? Gastroenterol. Hepatol. 2012, 8, 29–38. 7. Krzystek-Korpacka, M.; Kempi ´nski, R.; Bromke, M.; Neubauer, K. Biochemical biomarkers of mucosal healing for inflammatory bowel disease in adults. Diagnostics 2020, 10, 367. [CrossRef] 8. Bromke, M.A.; Neubauer, K.; Kempi ´nski, R.; Krzystek-Korpacka, M. Faecal calprotectin in assessment of mucosal healing in adults with inflammatory bowel disease: A meta-analysis. J. Clin. Med. 2021, 10, 2203. [CrossRef] 9. Nebbia, M.; Yassin, N.A.; Spinelli, A. Colorectal cancer in inflammatory bowel disease. Clin. Colon. Rectal Surg. 2020, 33, 305–317. [CrossRef] [PubMed] 10. Magro, F.; Rahier, J.F.; Abreu, C.; MacMahon, E.; Hart, A.; van der Woude, C.J.; Gordon, H.; Adamina, M.; Viget, N.; Vavricka, S.; et al. Inflammatory bowel disease management during the COVID-19 outbreak: The ten do's and don'ts from the ECCO-COVID Taskforce.
J. Crohns Colitis 2020, 14, S798–S806. [CrossRef] 11. Perisetti, A.; Goyal, H. Successful distancing: Telemedicine in gastroenterology and hepatology during the COVID-pandemic. Dig. Dis. Sci. 2021, 66, 945–953. [CrossRef] 12. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. JMLR 2011, 12, 2825–2830. 13. Seyed Tabib, N.S.; Madgwick, M.; Sudhakar, P.; Verstockt, B.; Korcsmaros, T.; Vermeire, S. Big data in IBD: Big progress for clinical practice. Gut 2020, 69, 1520–1532. [CrossRef] 14. Okagawa, Y.; Abe, S.; Yamada, M.; Oda, I.; Saito, Y. Artificial Intelligence in Endoscopy. Dig. Dis. Sci. 2021, 91, 1–20. [CrossRef]