2020 IEEE Student Conference on Research and Development (SCOReD)
27-28 September 2020, Johor, Malaysia

A Comparison of Machine Learning Approaches for Predicting the Progression of Crohn's Disease

Zain U. Hussain
College of Medicine and Veterinary Medicine
University of Edinburgh, UK
zain.hussain2@nhs.scot

Ragnor Comerford
College of Science and Engineering
University of Edinburgh, UK
research@ragnor.co

Fynn Comerford
College of Medicine and Veterinary Medicine
University of Edinburgh, UK
comerford.fynn@gmail.com

Nathan Ng
College of Medicine and Veterinary Medicine
University of Edinburgh, UK
nathanghuang@gmail.com

Dominic Ng
Faculty of Biology, Medicine and Health
. University of Manchester, UK
dominicmarkng1@gmail.com

Ateeb Khan
Barking, Havering and Redbridge University Hospitals NHS Trust, UK
makhan4395@gmail.com

Charlie Lees
College of Medicine and Veterinary Medicine
University of Edinburgh, UK
charlie.lees@ed.ac.uk

Amir Hussain
School of Computing
Edinburgh Napier University, UK
a.hussain@napier.ac.uk

***Abstract*—The incidence of Crohn's disease (CD) is rising, which calls for more accurate and less invasive diagnostic tools. The concentration of Faecal Calprotectin (FC) is a reliable indicator of luminal inflammatory processes and can replace invasive and uncomfortable ileocolonoscopies. Studies have confirmed the association of FC levels with the progression of CD and various machine learning approaches have been used for predicting disease progression. In this study, we aimed to comparatively evaluate the performance of established machine learning approaches, to predict the progression of CD, using a range of variables, including FC levels. Our dataset consisted of records for 804 patients with CD and a FC measurement, from a teaching hospital that cares for secondary and tertiary referred patients. We compared the performance of four machine learning approaches, namely logistic regression, support vector machine, random forests and artificial neural networks, to predict the likelihood of a flare up. Our results showed that all four approaches performed strongly, which demonstrates the potential of these approaches, in particular logistic regression, for predicting disease progression. Logistic regression slightly outperformed the others, with an accuracy of 0.90 and an AUC of 0.83. Our dataset had missing data for a number of patients, which resulted in fewer variables being selected for inclusion in the model. Our relatively small sample size could account for SVM, Random Forest and the ANN not demonstrating superior accuracy compared to logistic regression, in this study. In future, an increased number of variables should be included for analysis, the outcome period for a flare up should be explored, and our results should be validated using another independent and large dataset.**

***Keywords*—faecal calprotectin (FC), predictive modelling, irritable bowel disease (IBD), crohn's disease (CD), logistic regression**

I. INTRODUCTION

Crohn's disease (CD) is one of the two most prominent clinically defined types of Inflammatory bowel disease (IBD), which refers to chronic inflammatory disorders of the digestive tract culminating in accumulative damage of the digestive tract. CD, as well as Ulcerative Colitis, show periodic alternations between relapsing and remitting stages. While the exact cause of CD remains unclear, an early diagnosis enhances the success of subsequent therapies and is associated with fewer complications [1]. The absence of pathognomonic signs makes invasive procedures such as colonoscopies and/or histopathological examinations indispensable. These procedures are costly, risky and of considerable discomfort to the patient [2] [3]. Despite the rising incidence of CD [4], particularly in the pediatric population [5] [6], a large proportion of colonoscopies with suspected IBD show no pathological abnormalities [7]. This finding highlights the necessity for non-invasive and more discriminating diagnostic tests, to justify the need for endoscopic evaluations and facilitate early diagnoses of CD. The screening of faecal calprotectin (FC) levels has been shown as clinically useful. Increased concentrations of FC in patients' stool is a reliable marker of IBD and, hence, suitable for assessing disease progression in patients suffering from CD and for reducing the number of unnecessary endoscopies, as shown by van Rheenen et al. [8] from the University Medical Center Groningen [8]. FC is a dimeric calcium, iron, manganese and zinc sequestering protein [9] [10] [11]

978-1-7281-9317-5/20/$31.00 ©2020 Crown

and is the most abundant protein in neutrophils [12]. Although FC is also expressed at low levels in other phagocytic cells [12], it can be considered a neutrophil-specific biomarker of polymorphonuclear leukocyte infiltration of intestinal mucosa in this context [8]. The subsequent inflammatory process that most commonly affects the small intestine and the beginning of the colon in CD patients compromises the structural integrity of the mucosa. This results in neutrophil leakage into the lumen and their subsequent excretion with faeces is associated with a release of calprotectin which can be detected via enzyme-linked immunoabsorbent assays (ELISA). However, endoscopies are still necessary in patients of higher age for excluding more serious pathologies, such as cancer. Over the last decade, machine learning has been applied extensively in predicting disease progression, including a recent study which identified FC as a significant risk factor for predicting the progression of CD [13]. The best choice of machine learning models is not always apparent and requires comparative analyses to identify the model with optimal predictive performance. This approach is currently being been utilised for a range of clinical applications, including predicting asthma exacerbations [14]. Our study is the first of its kind to apply a comparative analysis of the performance of supervised machine learning methods for predicting the progression of CD, using FC levels. The aim of the predictive models is to discriminate between patients that are likely to reach the a primary composite endpoint (a flare up) and those that are not. The primary endpoint represents a composite of: a progression in montreal luminal behaviour, hospitalisation for flare up and resectional surgery.

II. DATA SET

This was a retrospective study of patients with a confirmed diagnosis of CD from the time period of 2003 to 2014. The patients were identified from a database at a teaching hospital that cares for secondary- and tertiary-referred patients with IBD. The primary inclusion criteria were a diagnosis of CD and at least 1 FC level measurement more than 3 months after diagnosis. During the study period, we identified 804 patients that fufilled the study inclusion criteria. Initial feature engineering and pre-processing was performed for the correct representation of categorical, ordinal and numerical data. Categorical features were encoded using one-hot encoding and numerical data normalized.

From the originally 58 features, we performed a ten-fold cross validation with a standard backward selection criteria as described in a study by Zeeshan et al., and extracted the most significant features from the input at a 95% confidence interval which can be found in Table I [15]. All analysis was carried out in Python, using the Scikit-learn library.

TABLE I. IDENTIFIED SIGNIFICANT PREDICTIVE FEATURES

Predictor	**Coeff.**	**S.Error**	**O. Ratio**	**z-value**	**p-value<=** 0.0001
Sex	1.15	.3065	3.168	3.719	0.0001
Montreal location at diagnosis	-0.19	0.1134	0.827	-1.58	0.0469
Harvey Bradshaw index at time of FC sample	-0.39	0.1755	0.675	-2.31	0.0104
Time since reassessment of Montreal behaviour to FC sample in months	-0.013	0.0057	0.987	-1.308	0.0954
Max Montreal behaviour	0.58	0.1669	1.791	3.642	0.0001
Platelet count	-0.003	0.0008	0.996	-3.52	0.000216
Abdominal Pain Score	0.56	0.3249	1.746	5.37	0.0001
Number of liquid stools per day	0.45	0.2047	1.563	2.1817	0.0145
Mouth Ulcers	2.01	0.0389	7.404	51.46	0.0001
FC count	0.33	0.0937	1.388	3.43	0.0003
Min FC	-0.01	0.009	0.996	-4.45	0.0001
Average FC	0.01	0.0016	1.008	4.98	0.0001
CompCount FC	-0.48	0.1284	0.617	-3.89	0.0001
CompMax FC	-0.01	0.0007	0.997	-4.29	0.0001
AvgComp FC	1.38	0.2978	4.009	4.662	0.0001
IncMaxNum	-0.01	0.0008	0.998	-2.50	0.00621
IncAvgCal	-0.94	0.3217	0.387	-2.95	0.0015

A. Logistic Regression

Logistic regression is a statistical model in which the probability of an outcome variable is approximated by applying the sigmoid function to a linear combination of potential predictor variables. This is equivalent to assuming a linear relationship between the predictor variables and the log-odds of the outcome. Mathematically, this can be formulated in the following form:

$$logit(E[Y_i \mid X_i]) = logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{\beta} \cdot \mathbf{X}_i \quad (1)$$

Model fitting is usually performed using maximum likelihood estimation for which currently no closed-form solution exists. However, it can be solved using an iterative process such as Newton's method or gradient descent.

B. Artificial Neural Networks

Artificial Neural Networks (ANN) are algorithms vaguely inspired by the structure of the human brain [16]. ANNs can be described as a set of nodes with activation functions connected by weighted direct links. ANNs are typically arranged in three layers (input, hidden, output), where the input layer represents our predictor variables and the output layer the predicted output of our network. The hidden nodes compute intermediate values and allow us to model complex non-linear relationships. In our study, we determine the topology of the neural network in terms of the size and number of hidden layers using hyperparameter optimisation and, finally, optimise the weights of the network using backpropagation of the error.

C. Random Forests

Random forests are an ensemble learning method that fit a set of decision tree classifiers on various sub-samples of the

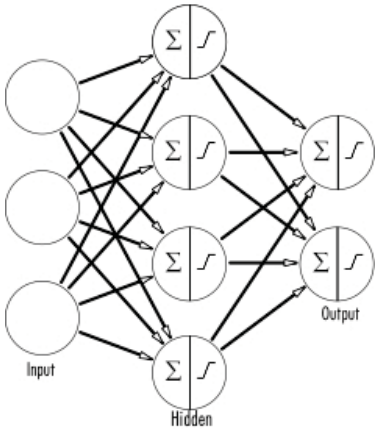

Fig. 1. Diagram of an artificial neural network

data set in parallel and uses the majority decision of the trees as the final decision. [17]. In contrast to individual high-variance decision trees, random forests are less easy to interpret but are able reduce bias and variance.

D. Support Vector Machine

The support vector machine (SVM) is one of the most popular supervised learning models. It constructs a maximum margin separating hyperplane that maximizes the distance to the nearest training-data point of any class. In the case that the original input space is not linearly separable, SVMs can make use of the so-called kernel trick and embed the data into a higher-dimensional space which is quite often easily separable.

III. EXPERIMENTAL RESULTS AND DISCUSSION

In order to avoid overfitting, we initially performed 10-fold cross validation to tune the regularization parameters of our four machine learning models.

The parameters of the different models, as seen in table II, were selected by their respective area under the Receiver Operating Characteristic curve (ROC), averaged over all validation sets. The ROC is created by plotting the true positive rate (TPR) against the false positive rate (FPR) as the discrimination threshold varies. The performance of the four models on the test set was then evaluated using the selected parameters from the validation set. The accuracy and area under the curve can be found in Table III.

We can observe that all models achieved comparable accuracy on the test set, with logistic regression demonstrating slightly superior accuracy (0.90). These results are consistent with the systematic review by Christodoulou et al. [16], which found machine learning to not outperform logistic regression in clinical prediction tasks. However, studies have shown machine learning to be superior in performance, in particular for large and complex datasets. It would be interesting to repeat our study with a larger dataset and an increased

TABLE II. IDENTIFIED HYPERPARAMETERS

Model	Parameter	Identified Optimal Value
Logistic Regression	penalization norm	l2
Logistic Regression	inverse of regularization strength	4.281
Neural Network	hidden layer size	(100, 75, 50)
Neural Network	activation function	logistic
Neural Network	solver	adam
Neural Network	l2 penalty	0.5
Neural Network	learning rate schedule	adaptive
Random Forest	number of trees	700
Random Forest	maximum depth	13
Random Forest	minimum number of samples required to split an internal node	2
Support Vector Machine	inverse of regularization strength	1
Support Vector Machine	kernel	linear

TABLE III. RESULTS

Model	**Accuracy**	**AUC**
Logistic Regression	0.9012	0.8285
Neural Network	0.88016	0.75
Random Forest	0.8802	0.7972
Support Vector Machine	0.9008	0.7931

number of variables, to test the hypothesis for this clinical prediction task. When looking at the Area Under the Curve (AUC), we can observe that logistic regression outperformed the other models. The random forest and support vector machine achieved similar scores, whereas the neural network architecture significantly underperformed.

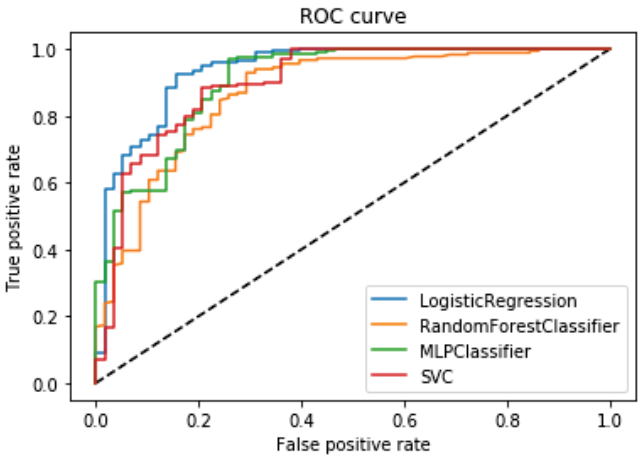

Fig. 2. ROC curve

IV. CONCLUSION

Taking into account the performance of the model, its complexity and interpretability, we can conclude that

logistic regression is the preferred model for predicting the progression of Crohn's disease using this dataset. Logistic regression had a slightly superior accuracy (0.90) compared to all three machine learning approaches, and its AUC was also noticeably superior (0.83). It should be noted that we selected only 17 features in our analysis as they resulted in the highest accuracy, and that some of the features that were excluded were due to missing data. The absence of an outcome period prediction in our analysis of the progression of Crohn's disease is a limitation. Further analysis should include the prediction of the time until a patient reaches the composite endpoint. Future studies should validate our results against other independent data sets, preferably with a larger sample size and more complete data, and should explore other secondary endpoints. The development of a clinically validated model for predicting the progression of CD would be an invaluable tool for clinicians to help predict flare ups and ensure early care interventions are provided.

REFERENCES

[1] R. Banerjee, M. Srikanth, B. R. R. Thanugundla, S. Dandaboyna, and N. D. Reddy, "Sa1233 early diagnosis of crohn's disease (cd) is associated with lesser complications," *Gastroenterology*, vol. 148, no. 4, pp. S–265, 2015.

[2] V. Lohsiriwat, "Colonoscopic perforation: incidence, risk factors, management and outcome," *World journal of gastroenterology: WJG*, vol. 16, no. 4, p. 425, 2010.

[3] S. L. Ristvedt, E. G. McFarland, L. B. Weinstock, and E. P. Thyssen, "Patient preferences for ct colonography, conventional colonoscopy, and bowel preparation," *The American journal of gastroenterology*, vol. 98, no. 3, pp. 578–585, 2003.

[4] C. S. Gismera and B. S. Aladrén, "Inflammatory bowel diseases: a disease (s) of modern times? is incidence still increasing?" *World Journal of Gastroenterology: WJG*, vol. 14, no. 36, p. 5491, 2008.

[5] H. M. Malaty, X. Fan, A. R. Opekun, C. Thibodeaux, and G. D. Ferry, "Rising incidence of inflammatory bowel disease among children: a 12-year study," *Journal of pediatric gastroenterology and nutrition*, vol. 50, no. 1, pp. 27–31, 2010.

[6] P. Henderson, R. Hansen, F. L. Cameron, K. Gerasimidis, P. Rogers, M. W. Bisset, E. L. Reynish, H. E. Drummond, N. H. Anderson, J. Van Limbergen *et al.*, "Rising incidence of pediatric inflammatory bowel disease in scotland," *Inflammatory bowel diseases*, vol. 18, no. 6, pp. 999–1005, 2012.

[7] A. Lasson, A. Kilander, and P.-o. Stotzer, "Diagnostic yield of colonoscopy based on symptoms," *Scandinavian journal of gastroenterology*, vol. 43, no. 3, pp. 356–362, 2008.

[8] P. F. Van Rheenen, E. Van de Vijver, and V. Fidler, "Faecal calprotectin for screening of patients with suspected inflammatory bowel disease: diagnostic meta-analysis," *Bmj*, vol. 341, p. c3369, 2010.

[9] J. Z. Liu, S. Jellbauer, A. J. Poe, V. Ton, M. Pesciaroli, T. E. Kehl-Fie, N. A. Restrepo, M. P. Hosking, R. A. Edwards, A. Battistoni *et al.*, "Zinc sequestration by the neutrophil protein calprotectin enhances salmonella growth in the inflamed gut," *Cell host & microbe*, vol. 11, no. 3, pp. 227–239, 2012.

[10] T. G. Nakashige, B. Zhang, C. Krebs, and E. M. Nolan, "Human calprotectin is an iron-sequestering host-defense protein," *Nature chemical biology*, vol. 11, no. 10, pp. 765–771, 2015.

[11] S. M. Damo, T. E. Kehl-Fie, N. Sugitani, M. E. Holt, S. Rathi, W. J. Murphy, Y. Zhang, C. Betz, L. Hench, G. Fritz *et al.*, "Molecular basis for manganese sequestration by calprotectin and roles in the innate immune response to invading bacterial pathogens," *Proceedings of the National Academy of Sciences*, vol. 110, no. 10, pp. 3841–3846, 2013.

[12] A. Røseth, P. Schmidt, and M. Fagerhol, "Correlation between faecal

excretion of indium-111-labelled granulocytes and calprotectin, a granulocyte marker protein, in patients with inflammatory bowel disease," *Scandinavian journal of gastroenterology*, vol. 34, no. 1, pp. 50–54, 1999.

[13] N. A. Kennedy, G.-R. Jones, N. Plevris, R. Patenden, I. D. Arnott, and C. W. Lees, "Association between level of fecal calprotectin and progression of crohn's disease," *Clinical Gastroenterology and Hepatology*, vol. 17, no. 11, pp. 2269–2276, 2019.

[14] Z. Hussain, S. A. Shah, M. Mukherjee, and A. Sheikh, "Predicting the risk of asthma attacks in children, adolescents and adults: protocol for a machine learning algorithm derived from a primary care-based retrospective cohort," *BMJ open*, vol. 10, no. 7, p. e036099, 2020.

[15] Z. K. Malik, Z. U. Hussain, Z. Kobti, C. W. Lees, N. Howard, and A. Hussain, "A new recurrent neural network based predictive model for faecal calprotectin analysis: A retrospective study," *arXiv preprint arXiv:1612.05794*, 2016.

[16] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *Journal of clinical epidemiology*, vol. 110, pp. 12–22, 2019.

[17] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.