2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

mRNA Robust Signature for IBD Using Machine Learning

1st David Rojas-Velazquez
Division of Pharmacology, University of Utrecht, Department of Data Science,
Julius Center for Health Sciences *and Primary Care, University Medical Center Utrech The Netherlands*
e.d.rojasvelazquez@uu.nl

2nd Sarah Kidwai
Division of Pharmacology, University of Utrecht
Utrecht, The Netherlands
s.kidwai@uu.nl

3rd Luciënne de Vries
Division of Pharmacology, University of Utrecht
Utrecht, The Netherlands
l.devries14@students.uu.nl

4th Johan Garssen
Division of Pharmacology, University of Utrecht, Global Centre of Excellence Immunology Danone Nutricia Research
Utrecht, The Netherlands
J.Garssen@uu.nl

5th Alberto Tonda
UMR 518 MIAPS, INRAE, Université Paris-Saclay, Institut des Systèmes Complexes de Paris ÎledeFrance (ISC-PIF) - UAR 3611 CNRS
Paris, France
alberto.tonda@inrae.fr

6th Alejandro Lopez-Rincon
Division of Pharmacology, University of Utrecht, Department of Data Science, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht The Netherlands
a.lopezrincon@uu.nl

***Abstract*—Inflammatory bowel disease, including Crohn's disease and ulcerative colitis, is a rising global issue. Accurate diagnosis is vital but challenging. This study used the REFS algorithm to identify IBD biomarkers using three mRNA datasets from the GEO repository. The selected genes demonstrated excellent diagnostic accuracy, highlighting the potential of machine learning in advancing IBD research.**

***Index Terms*—REFS, biomarkers, IBD, mRNA processing, bioinformatics**

I. Introduction

Inflammatory bowel disease (IBD) is a rising global health issue. It includes Crohn's disease, which can cause inflammation anywhere in the gastrointestinal tract, and ulcerative colitis, which only affects the colon's mucosal layer [1]. Accurate diagnosis is difficult due to various clinical factors. Delayed or incorrect diagnosis can worsen the disease and complicate remission [2]. New high-throughput technologies allow omics data analysis, using bioinformatics tools and machine learning methods, to achieve precision medicine goals. These tools are promising for identifying clinically relevant patterns and predictive markers for complex diseases like IBD [2]. In this study, we used the REFS algorithm [4], [5] on three mRNA datasets from the GEO repository to identify potential IBD biomarkers.

II. Matherials and Methods

A. mRNA datasets

The three datasets used in this study were downloaded from the Gene Expression Omnibus (GEO) repository: 1) accession number GSE3365 [3] (127 samples: 85 IBD / 42 controls), 2) accession number GSE71730 [6] (47 samples: 37 IBD / 10 controls), and 3) accession number GSE33943 [7] (58 samples: 45 IBD / 13 controls). Samples were categorized into two groups: control (0) and IBD (1), which includes CD and UC patients. The GSE3365 dataset, having the most samples, was chosen for discovery.

B. REFS

The Recursive Ensemble Feature Selection (REFS) algorithm is used to identify biomarkers by determining which features are most effective in distinguishing between case and control groups [4], [5]. The ensemble is composed by 8 classifiers from the scikit-learn toolbox [8]: Stochastic Gradient Descent (SGD) on linear models, Support Vector Machine classifier (SVC), Gradient Boosting, Random Forest, Logistic Regression, Passive Aggressive classifier, Ridge Classifier and Bagging. REFS uses a nested-cross validation within a 10-fold scheme, ensuring accurate and unbiased results [9], and a validation module to minimize gene selection bias with five additional classifiers from the scikit-learn toolkit [8]: AdaBoost, Extra Trees, KNeighbors, MLP, and LassoCV. The AUC metric, derived from the average accuracy of five classifiers in a nested 10-fold cross-validation, measures the effectiveness of a discriminant test. Near 1.0 values signify high performance [10].

III. Results

Using the GSE3365 dataset, REFS identified 16 out of 22,283 genes as the most effective for distinguishing IBD patients. With these genes, REFS achieved its highest accuracy

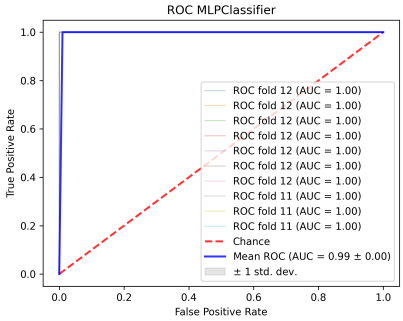

Fig. 1. Individual AUC for the classifier with the best performance in the validation process (MLP) was calculated using the 16 genes selected by REFS from GSE3365.

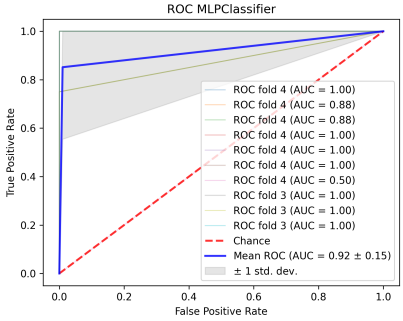

Fig. 2. Individual AUC for the classifier with the best performance in the validation process (MLP) was calculated using the 16 genes selected by REFS in the GSE71730 test database.

(over 0.97) in its feature selection module. The validation module in REFS applied to the 16 selected genes yielded an average AUC of 0.99, the classifier with the best performance was the Multi-Layer Perceptron (MLP) with an individual AUC of 0.99, see Fig. 1. The resulting average AUC corresponds to a "*excellent*" diagnostic accuracy [10]. The efficiency of the 16 genes selected by REFS was compared against SelectKBest algorithm ($k = 16$) from the scikit-learn toolbox [8], and a 10-times repeated random selection. The average AUCs were 0.946 for SelectKBest and 0.7194 for random selection. The genes with higher AUCs chosen by REFS were then tested on two datasets: GSE71730 and GSE33943.

After identifying the 16 genes selected by using REFS in GSE71730 and GSE33943 testing datasets, REFS validation yielded an average AUC of 0.874 and 0.894 respectively. Both AUC correspond to "*very good*" diagnostic accuracy [10]. The classifiers with the best performance were MLP with an individual AUC of 0.92 for GSE71730, see Fig. 2, and LassoCV with an individual AUC of 0.90 for GSE33943, see Fig. 3. A 10-time random validation was done on both testing datasets, randomly selecting 16 genes each time. REFS validation was applied, the resulting average AUCs for GSE71730 was 0.66 and 0.7472 for GSE33943. Finally, after identifying the 16 genes selected by using SelectKBest in GSE71730 and GSE33943 testing datasets, REFS validation yielded an average AUC of 0.784 and 0.862 respectively.

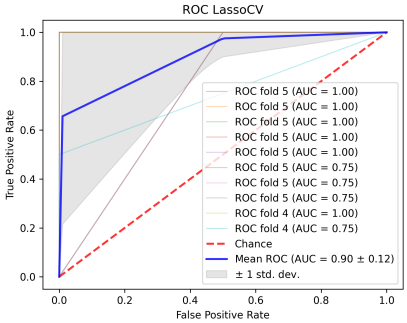

Fig. 3. Individual AUC for the classifier with the best performance in the validation process LassoCV was calculated using the 16 genes selected by REFS in the GSE33943 test database.

IV. Conclusion

This study used the REFS algorithm to identify potential IBD biomarkers from three mRNA databases. One database was used for discovery and the others for testing. The selected genes by using REFS showed better diagnostic accuracy compared by those selected by using SelectKBest and 10-time random validation. Despite limited information on these genes' impact on IBD, the approach can yield promising results for IBD diagnosis and treatment development.

References

[1] Y. Z. Zhang, and Y. Y. Li, "Inflammatory bowel disease: pathogenesis," World journal of gastroenterology: WJG, 20(1), 2014, pp. 91–99. doi.org/10.3748/wjg.v20.i1.91.

[2] B. Stankovic, N. Kotur, G. Nikcevic, V. Gasic, B. Zukic, and S. Pavlovic, "Machine learning modeling from omics data as prospective tool for improvement of inflammatory bowel disease diagnosis and clinical classifications," Genes, 12(9), 2021, p. 1438. doi: 10.3390/genes12091438.

[3] M.E. Burczynski, et al., "Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells," The journal of molecular diagnostics, 8(1), 2006, pp. 51–61. doi.org/10.2353/jmoldx.2006.050079.

[4] A. Lopez-Rincon, M. Martinez-Archundia, G. U. Martinez-Ruiz, A. Schoenhuth, and A. Tonda, "Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection," BMC bioinformatics, 20, 2019, pp. 1–17. doi.org/10.1186/s12859-019-3050-8.

[5] K. Kamphorst, A. Lopez-Rincon, A. M. Vlieger, J. Garssen, E. van't Riet, and R. M. van Elburg, "Predictive factors for allergy at 4–6 years of age based on machine learning: A pilot study," PharmaNutrition, 23, 2023, p. 100326. doi.org/10.1016/j.phanu.2022.100326.

[6] B. Gurram, et al., "Plasma-induced signatures reveal an extracellular milieu possessing an immunoregulatory bias in treatment-naive paediatric inflammatory bowel disease," Clinical and Experimental Immunology 184(1), 2016, pp. 36–49. doi.org/10.1111/cei.12753.

[7] P. P. E. van Lierop, et al., "Gene expression analysis of peripheral cells for subclassification of pediatric inflammatory bowel disease in remission," PLoS One 8(11), 2013, p. e79549. doi.org/10.1371/journal.pone.0079549.

[8] F.Pedregosa, et al., "Scikit-learn: Machine learning in Python," the Journal of machine Learning research, 12, 2011, pp. 2825-2830.

[9] A. Vabalas, et al, "Machine learning algorithm validation with a limited sample size," PloS one 14(11), 2019, p. e0224365. doi.org/10.1371/journal.pone.0224365.

[10] A. M. Šimundić, "Measures of diagnostic accuracy: basic definitions," ejifcc, 19(4), 2009, p. 203.