Machine-Learning Analysis of mRNA: An Application to Inflammatory Bowel Disease

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
Utrecht, 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

4rd Péter Tözsér
Division of Pharmacology,
University of Utrecht
Utrecht, The Netherlands
p.tozser@students.uu.nl

5th Luis Oswaldo Valencia-Rosado
Department of Computing, Electronics and
Mechatronics, Universidad de las Américas Puebla
San Andrés Cholula, Mexico
luis.valencia@udlap.mx

6th Johan Garssen
Division of Pharmacology,
University of Utrecht,
Global Centre of Excellence
Immunology Danone Nutricia Research
Utrecht, The Netherlands
J.Garssen@uu.nl

7th Alberto Tonda
UMR 518 MIA-PS, INRAE,
Université Paris-Saclay, Institut
des Systèmes Complexes de Paris
Île-de-France (ISC-PIF)
- UAR 3611 CNRS
Paris, France
alberto.tonda@inrae.fr

8th 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 (IBD), that includes Crohn's disease (CD) and Ulcerative Colitis (UC), is a global health concern due to the increasing number of cases. Diagnosing IBD is a challenging task due to a considerable number of clinical factors. Delayed or inaccurate IBD diagnosis can worsen the disease and complicate achieving remission, therefore, early diagnosis and prompt treatment are crucial. In this study, we adapted a methodology to analyze 16s rRNA (18,758 features) to analyze mRNA (54,675 features) that consists of three phases: 1) preprocessing, 2) feature selection, and 3) testing. We applied this methodology for analyzing mRNA datasets from the Gene Expression Omnibus (GEO) repository, aiming to discover possible biomarkers for IBD diagnosis. We experimented with three datasets, using one dataset for feature (gene) selection and we tested the results in the other two. We compared results with those obtained from other feature selection methods, such as the F-score-based K-Best and random selection. The Area Under the Curve (AUC) was used to measure the diagnostic accuracy and as a metric to compare results between the methodology and other feature selection methods. The Matthews Correlation Coefficient (MCC) was used as an additional metric to evaluate the performance of the methodology and for comparison with other feature selection methods.**

***Index Terms*—REFS, biomarkers discovery, mRNA processing**

I. Introduction

Inflammatory Bowel Disease (IBD) comprises two separate chronic conditions that cause inflammation in the gastrointestinal tract: Crohn's disease (CD) and Ulcerative Colitis (UC) [1]. CD may cause inflammation in the entire wall of the gastrointestinal tract, typically found in the terminal ileum or in the perianal region [1].On the other hand, UC is manifested by inflammation that affects only the mucosal layer of the colon [1]. A late or inaccurate diagnosis of IBD can negatively impact disease progression, often causing complications that make it difficult to achieve or maintain remission [2]. Accurately diagnosing IBD is challenging because it relies on the analysis of a combination of clinical studies, including endoscopy or colonoscopy, stools and blood samples [2]. Rapid and accurate diagnosis of IBD still depends on the discovery of reliable biomarkers and their integration into clinical practice, which is crucial to significantly improve patient outcomes. [2]. Thus it is necessary to find a proper technique for biomarker discovery such as ML.

The use of new high-throughput technologies has enabled

the analysis of omics data, including mRNA, miRNA, protein, DNA methylation, and histone modifications. This analysis combines bioinformatics tools and Machine Learning (ML) algorithms to find new clinical patterns and reliable predictive biomarkers for IBD [2]. In recent years, mRNA gene expression profiles have been used to find potential biomarkers for IBD diagnosis. Several studies employ machine learning tools to analyze mRNA from various samples, such as blood [3]–[6], mucosal biopsies [7]–[10] and intestinal biopsies [11]–[14] for diagnosing IBD. Despite the promising results, there is a need to improve the identification of reliable biomarkers in IBD patients to improve diagnostic procedures [3].

In this study, We adapted a methodology to analyze 16s rRNA (18,75 features) [43] to analyze mRNA (54,675 features) datasets obtained from the Gene Expression Omnibus (GEO) to identify possible biomarkers to diagnose IBD. To the best of our knowledge, there are no research works that adapt the aforementioned methodology to analyze other types of omics data. The methodology consists of three phases:

- **Preprocessing:** raw data analysis and/or series matrix analysis.
- **Feature selection:** the Recursive Ensemble Feature Selection (REFS) algorithm for biomarker discovery [15], [16] is applied to one dataset selected as *"discovery dataset"* to identify possible biomarkers.
- **Testing:** test the effectiveness of the feature set selected by using REFS in at least two different datasets.

II. Materials and Methods

A. mRNA datasets

The three datasets analyzed were downloaded from the Gene Expression Omnibus (GEO) repository. The selection criteria was based on the *"dataset selection criteria"* defined in [43]: 1) mRNA extracted from the same source (blood samples), 2) datasets containing both case and control groups, and 3) at least 10 samples in each group. The three selected datasets were generated in different laboratories using different technologies:

- **Accession number GSE3365 [4]:** Peripheral Blood Mononuclear Cells (PBMC) with 127 samples, where 85 samples belong to the IBD group and 42 samples belong to the control group.
- **Accession number GSE71730 [17]:** Blood plasma with 47 samples, where 37 samples belong to the IBD group and 10 samples belong to the control group.
- **Accession number GSE33943 [18]:** Peripheral Blood Leucocytes (PBL) with 58 samples, where 45 samples belong to the IBD group and 13 samples belong to control group.

Samples were labeled as 0 for the control group, including people not diagnosed with IBD or patients in remission, and 1 for the IBD group, including patients with CD, UC, or both. The dataset with accession number GSE3365 was selected as the discovery dataset since it has the largest number of samples.

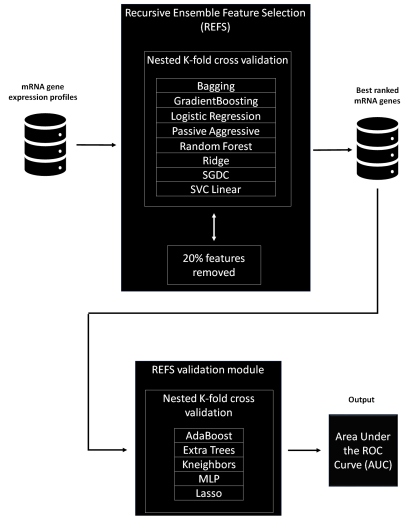

Fig. 1. REFS workflow and validation module workflow overviews.

B. REFS

The Recursive Ensemble Feature Selection (REFS) algorithm is used to identify potential biomarkers by selecting the features that are most effective for a discriminative analysis with the highest accuracy using the minimum number of features [15]. This algorithm has been applied in numerous studies involving the analysis of omics data, including miRNA [15], [19], mRNA [20]–[23], and 16s rRNA [16], [24]–[27], obtaining promising results. In addition, some results have been validated in laboratory tests for SARS-CoV-2 detection [46], [47]. The ensemble is made up of eight classifiers from the scikit-learn toolbox [28]: Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), Gradient Boosting (GB), Random Forest (RF), Logistic Regression (LR), Passive Aggressive classifier (PAc), Ridge Classifier (RC) and Bagging.

Working with high-dimensional and low-sample-size datasets can be a real challenge because it can lead models to overfitting and biased results [29]. To prevent this issue, REFS implements nested strategy in the 10-fold cross-validation setup to generate results that are more precise and free from bias, even when the sample size is relatively small, for detailed nested startegy see [29]. REFS removes 20% of the least important features (genes) each time until only one remains. This procedure is carried out concurrently at least 10 times to offset the randomness inherent in some classifiers in the ensemble (e.g., RF) and the 10-fold cross-validation process. Performance metrics were calculated for each run, and the best-performing set of features is selected.

To reduce bias selection, REFS employs a validation module that applies five classifiers from the scikit-learn toolkit [28] that are not part of the ensemble: AdaBoost, Extra Trees, KNeighbors, Multi-Layer Perceptron (MLP), and Least Absolute Shrinkage and Selection Operator plus an iterative process using Cross-Validation (LassoCV). Fig. 1 illustrates how the

REFS process is executed.

C. Experimental design

The experimental design in this study is based on the proposed methodology in [43]. This methodology is for reproducible 16s rRNA data analysis. The analysis is performed in three different datasets, one of which is selected as discovery while the resting two are saved for testing [43]. To validate the performance, the results are compared with another feature selection methods using Area Under the Curve (AUC) and Matthews Correlation Coefficient (MCC) as metrics to measure and compare performance [43].

In contrast with [43], we replaced the raw data processing with a *"Prepossessing phase"*, and made small changes in the dataset selection criteria such as selecting mRNA datasets instead of 16s rRNA datasets. On the other hand, we also made a comparison between the methodology proposed in this manuscript and other feature selection methods, using the same metrics for evaluating performance and comparison between methods.

After the corresponding modifications to the aforementioned methodology, we propose an experimental methodology to analyze mRNA datasets that consists of three phases, as illustrated in Fig. 2:

1) ***Preprocessing:*** this phase consists in extracting the necessary information from all datasets to be used in the feature selection and testing phases. This information is available in the "series matrix file" and includes labels for each group, gene reference IDs, and abundance data for each sample. This information will serve as input for the next phase.
2) ***Feature selection:*** this phase consists in executing REFS on the discovery dataset to identify the features most effective in differentiating between the control and IBD groups with the highest accuracy using the minimum number of features. Once the features are selected, we execute the validation module to test the performance of the feature set selected by REFS.
3) ***Testing phase:*** consists in searching for the feature set selected by REFS in at least two different datasets and extracting the abundance data corresponding to each sample in the datasets (testing datasets). Then, execute the validation module on each testing dataset, the inputs are labels extracted from the "series matrix file", reference ID, and abundance data extracted on each dataset.

The average accuracy of the five classifiers in the validation module provides a value for the AUC, which is the metric to evaluate REFS performance. Values close to 1.0 indicate excellent performance [30]. In the same way, the average accuracy given by the validation module provides a value for the Matthews Correlation Coefficient (MCC) where values close to -1 indicate poor accuracy and values close to 1 represent excellent accuracy [42]. For further validation, the effectiveness of REFS is compared against other feature selection techniques:

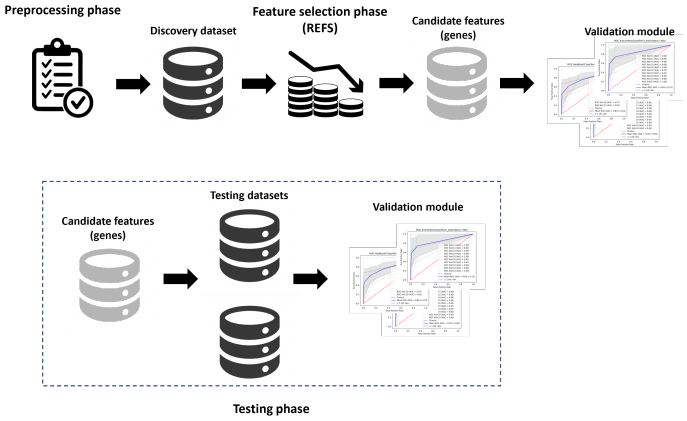

Fig. 2. Summary of the methodology. In contrast to other methodologies, we select the features in one dataset and we test in 2 other datasets. The workflow of the preprocessing and feature selection phases is shown above. The workflow of the testing phase is shown below.

- *K-Best with F-score*, implementing the SelectKBest algorithm from the scikit-learn toolbox [28], which is an algorithm that chooses the K top-scoring features.
- *10-time random feature selection*, consisting of randomly selecting N genes from the total genes in the datasets each time.

III. Results

A. Preprocessing phase

After analyzing the information within each "series matrix file", we extracted the total number of samples, identified the groups along with the number of samples in each, the total number of gene reference IDs (total features), and the counts of each gene in each sample, see Table I. According to this analysis, we selected GSE3365 as discovery because it contains the largest number of samples.

B. Feature selection phase

After applying REFS to the discovery dataset, 16 genes were selected as the most effective feature set for distinguishing IBD patients from the control group. With these 16 genes, REFS achieved its highest accuracy (above 0.9700), see Fig. 3a. After applying the validation module to the 16 selected genes, an average AUC of 0.9900 was obtained, as detailed in Table II. The classifier with the best performance was the Multi-Layer Perceptron (MLP) with an individual AUC of

TABLE I
CHARACTERISTICS OF THE SELECTED DATASETS

Accession Number	Samples	Labels	Gene Reference IDs
GSE3365*	127	85 - IBD 42 - control	22,283
GSE71730	47	37 - IBD 10 - control	54,675
GSE33943	58	45 - IBD 13 - control	54,675

* Discovery dataset.

TABLE II
RESULTS OF THE VALIDATION MODULE APPLIED TO THE DISCOVERY DATASET (GSE3365). THE STANDARD DEVIATION WAS OMITTED TO MAINTAIN SIMPLICITY AND AVOID COMPLEXITY IN THE TABLE

Classifier	REFS		SelectKBest		10-time random selection	
	AUC	*MCC*	*AUC*	*MCC*	*AUC*	*MCC*
AdaBoost	0.9900	0.9822	0.9200	0.8474	0.7050	0.4503
Extra Trees	0.9900	0.9900	0.9300	0.8658	0.7390	0.5264
KNeighbors	0.9900	0.9900	0.9700	0.8919	0.7220	0.4463
MLP	0.9900	0.9900	0.9600	0.9239	0.7450	0.5123
Lasso	0.9900	0.9900	0.9500	0.9508	0.6860	0.4200
Average	**0.9900**	**0.9884**	**0.9460**	**0.8959**	**0.7194**	**0.4711**

TABLE III
RESULTS OF THE VALIDATION MODULE APPLIED TO THE GSE71730 TESTING DATASET. THE STANDARD DEVIATION WAS OMITTED TO MAINTAIN SIMPLICITY AND AVOID COMPLEXITY IN THE TABLE

Classifier	REFS		SelectKbest		10-time random selection	
	AUC	*MCC*	*AUC*	*MCC*	*AUC*	*MCC*
AdaBoost	0.8300	0.6362	0.6200	0.5587	0.7170	0.4296
Extra Trees	0.9000	0.8000	0.8800	0.7612	0.6860	0.3897
KNeighbors	0.8900	0.7612	0.8700	0.7362	0.6270	0.2516
MLP	0.9200	0.7362	0.8800	0.7612	0.7460	0.5024
Lasso	0.8300	0.5612	0.6700	0.3612	0.5240	0.0435
Average	**0.8740**	**0.6989**	**0.7840**	**0.6357**	**0.6600**	**0.3234**

0.9900, see Fig. 3b. The resulting average AUC corresponds to a "*excellent*" diagnostic accuracy [30].

To compare the efficiency of the 16 genes selected by REFS, we compared its performance against two other feature selection methods: K-Best with ($k = 16$) and random selection with ($N = 16$). After performing the validation module, we obtained an average AUC of 0.9460 for the SelectKBest algorithm and 0.7194 for the random selection, see Table II. Using MCC as an additional metric to evaluate performance, REFS achieved better performance compared to the other feature selection methods with an average MCC of 0.9884, see Table II.

C. Testing phase

After identifying the 16 genes and their the abundance in the two testing datasets, we executed the validation module obtaining an average AUCs of 0.8740 for GSE71730 and 0.8940 for GSE33943, see Table III for GSE71730 and Table IV for GSE33943. The resulting average AUCs for both testing datasets corresponds to a "*very good*" diagnostic accuracy [30]. The classifier with the best performance for GSE71730 was Multi-Layer Perceptron (MLP), with an individual AUC of 0.9200, see Fig. 3c. The classifier with the best performance for GSE33943 was LassoCV with an individual AUC of 0.9000, see Fig. 3d. Performing the MCC, we obtained an average MCC of 0.6989 for GSE71730 and 0.6903 for GSE33943, see Table III for GSE71730 and Table IV for GSE33943.

To compare the efficiency of the 16 genes selected by REFS, we searched for the 16 genes selected by K-Best on each testing dataset. After executing the validation module, we obtained

TABLE IV
RESULTS OF THE VALIDATION MODULE APPLIED TO THE GSE33943 TESTING DATASET. THE STANDARD DEVIATION WAS OMITTED TO MAINTAIN SIMPLICITY AND AVOID COMPLEXITY IN THE TABLE

Classifier	REFS		SelectKbest		10-time random selection	
	AUC	*MCC*	*AUC*	*MCC*	*AUC*	*MCC*
AdaBoost	0.9000	0.7952	0.8100	0.6099	0.7160	0.4207
Extra Trees	0.8800	0.6447	0.8400	0.6632	0.7290	0.4985
KNeighbors	0.9000	0.8447	0.8400	0.6182	0.7980	0.5543
MLP	0.8900	0.5795	0.9100	0.7632	0.8410	0.6284
Lasso	0.9000	0.5877	0.9100	0.7010	0.6520	0.3552
Average	**0.8940**	**0.6903**	**0.8620**	**0.6711**	**0.7472**	**0.4914**

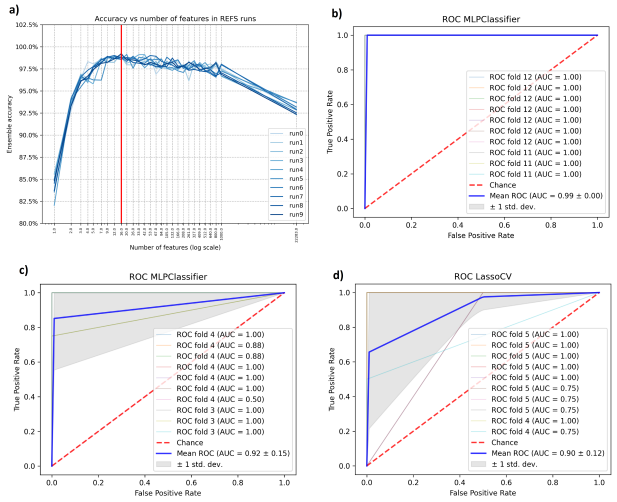

Fig. 3. a) The minimum number of features that achieve the highest accuracy, b) The classifier with the best performance in the discovery dataset GSE3365, c) The classifier with the best performance for GSE71730, and d) The classifier with the best performance for GSE33943. The shape of the ROC curves are explained by the binary outputs of the classifiers.

the average AUCs of 0.7840 for GSE71730 and 0.8620 for GSE33943. For the 10-time random selection, the resulting average AUCs were 0.6600 for GSE71730 and 0.7472 for GSE33943. Using the MCC as an additional metric to evaluate and compare performance, REFS achieved better performance compared with the other feature selection methods. The results are presented in detail in Table III for GSE71730 and Table IV for GSE33943.

IV. DISCUSSION

After executing the proposed methodology, 16 genes selected by REFS were sufficient to distinguish between both control and IBD groups. The performance evaluation, using different metrics including the Area Under the Curve (AUC) and Matthews Correlation Coefficient (MCC), shows that the proposed methodology has better performance compared with K-Best and random selection. The similar AUC and MCC values between REFS and K-Best can be explained due to the overlap in some genes selected by both algorithms (201032_at, 200867_at, 200680_x_at, and 212808_at). The 16 genes, iden-

TABLE V
IMPACT OF EACH GENE SELECTED BY USING REFS, UP- OR DOWN-REGULATED IN IBD PATIENTS.

ID_REF	GEO profile	Impact on IBD
201032_at	Bladder Cancer Associated Protein (BLCAP)	Down-regulated
200867_at	Ring Finger Protein 114 (RNF114)	Down-regulated
200680_x_at	High Mobility Group Box 1 (HMGB1)	Down-regulated
218999_at	Transmembrane Protein 140 (TMEM140)	Up-regulated
208546_x_at	Histone Cluster 1, H2bh (HIST1H2BH)	Up-regulated
213737_x_at	Golgin A8 Family Member N (GOLGA8N)	Down-regulated
209141_at	Ubiquitin Conjugating Enzyme E2 G1 (UBE2G1)	Down-regulated
202708_s_at	Histone Cluster 2, H2be (HIST2H2BE)	Up-regulated
218822_s_at	UIncharacterized LOC100652930 (LOC100652930)	Down-regulated
212808_at	Nuclear Factor of Activated T-cells 2 Interacting Protein (NFATC2IP)	Down-regulated
218798_at	KRI1 Homolog (KRI1)	Down-regulated
220143_x_at	LUC7 Like (LUC7L)	Down-regulated
211967_at	Transmembrane Protein 123 (TMEM123)	Up-regulated
202362_at	RAP1A, Member of RAS Oncogene Family (RAP1A)	Up-regulated
219983_at	HRAS Like Suppressor (HRASLS)	Up-regulated
201169_s_at	Basic Helix-loop-helix Family Member e40 (BHLHE40)	Down-regulated

tified by their ID_REF, along with their GEO profile [1] and their impact on IBD (whether up- or down-regulated), are presented in Table V. For the visual representation of the impact on IBD (whether up- or down-regulated) see Fig. 4.

It is important to mention, although the preprocessing phase described in this work is aimed at extracting information from the "series matrix file(s)", the processing of raw files (read counts) is also considered as part of the methodology for mRNA analysis. On the other hand, we keep the general approach applied in [43] where the reproducibility of experiments, as well as testing in different datasets, are important objectives in the area of biomarker discovery.

[1]https://www.ncbi.nlm.nih.gov/geo

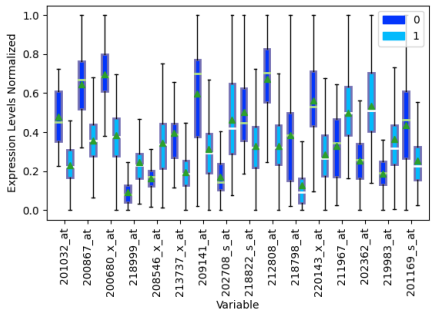

Fig. 4. Expression data of each feature in IBD patients (1) and the control group (0) in the GSE3365 discovery dataset.

Analyzing the biological information related to the 16 genes, we found that some of them have already been identified as biomarkers for IBD. For example:

- **BLCAP:** is down-regulated in IBD patients, this is consistent with previous research, where BLCAP was identified as a diagnostic marker for IBD. This gene regulates the proliferation of cells commonly found in intestinal tissue. Additionally, a slight increase was found in bladder cancer risk among IBD patients, but it was not enough to be statistically significant. [31]–[34].
- **RNF114:** is down-regulated in IBD patients. This gene functions as a regulator of the NF-κB pathway. Through ubiquitinating and stabilizing NF-κB signaling inhibitors A20 and IκBa, RNF114 can mildly decrease inflammation. A20 and IκBa inhibit the pathway at multiple points, making them potential drug targets for reducing inflammation in IBD patients [35]–[37].
- **HMGB1:** is a pro-inflammatory protein that is released from necrotic cells, activating multiple pro-inflammatory pathways. It has been found up-regulated in stool samples of IBD patients, with its levels correlating with disease severity. The difference in the behavior of this gene may be explained by the fact that our analysis is based on blood samples, while research on HMGB1 has been conducted using stool samples. HMGB1 has also been studied in relation to neurodegenerative diseases [38]–[40].
- **UBE2G1:** is a known biomarker for IBD, found to be down-regulated in IBD patients. Its function is to degrade short-lived proteins. Disturbed functionality of UBE2G1 has been linked to disease implications. Additionally, ubiquitin restricts inflammation by binding to A20 in the NF-κB pathway [41].

Not only can this methodology be used for diagnostic purposes, it can also be used to predict responses to medical treatments, for example, patient response to Infliximab therapy in where the mucosal mRNA data from [44] were analyzed using the same approach. The dataset consists of 61 mucosal biopsies from 61 IBD patients including untreated CD and UC samples. REFS selected, with an average AUC of 0.9840, nine genes from the original 54,675. The nine genes, identified by their ID_REF, along with their GEO profile and their impact on responders to the medical treatment (whether up- or down-regulated), are presented in Table VI. For visual representation of the impact on IBD (whether up- or down-regulated) see Fig 5.

The selected genes were searched in an independent database [45] to validate the effectiveness of our methodology in predicting therapy response. The same nine genes were found in the independent dataset. After performing the validation module, the average AUC of 0.6660 was achieved. These results motivate us to continue doing more research and experimentation with the proposed methodology applied to mRNA datasets for potential biomarker discovery.

Given the performance achieved by REFS, we can say

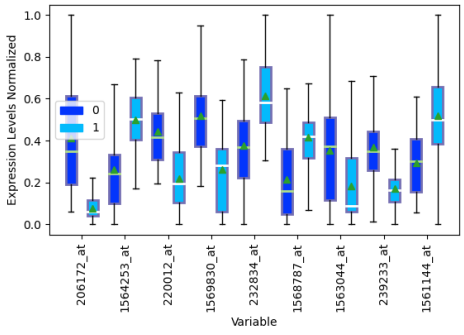

Fig. 5. Expression data of each feature in responders group (1) and the non-responders group (0) in the discovery dataset [44].

that the absence of biological information for the other genes suggests they may potentially serve as biomarkers for diagnosing IBD and for predicting patient response. Therefore more research should be done about these new findings. Although the datasets were generated by different laboratories, the results point to a generalized solution worth studying further.

V. Conclusion

We proposed a methodology to analyze mRNA, based on a 16s rRNA methodology that aims to solve the reproducibility problem identified in the biomarker discovery area and achieved the objective of testing results in different datasets [43]. This proposed methodology uses the Recursive Ensemble Feature Selection (REFS) algorithm to identify potential biomarkers for the diagnosis of Inflammatory Bowel Disease (IBD).

We analyzed three mRNA datasets obtained from the GEO repository, with one dataset used for discovery (GSE3365) and the other two (GSE71730 and GSE33943) for testing the results. Our methodology demonstrates better performance in the Area Under the Curve (AUC) and Matthews Correlation Coefficient (MCC) metrics compared to K-Best and 10-time random selection. The analysis of biological information of the 16 genes selected by REFS revealed consistency with established knowledge in this field.

TABLE VI
Impact of each gene selected by using REFS, Up- or Down-regulated in responders.

ID_REF	GEO profile	Impact on IBD
206172_at	Interleukin 13 Receptor Subunit Alpha 2 (IL13RA2)	Down-regulated
220012_at	Endoplasmic Reticulum Oxidoreductase 1 Beta (ERO1-L(BETA))	Up-regulated
1569830_at	Protein Tyrosine Phosphatase, Receptor Type C (PTPRC)	Down-regulated
239233_at	Coiled-Coil Domain Containing 88A (CCDC88A)	Down-regulated
1564253_at	Uncharacterized	Up-regulated
232834_at	Uncharacterized	Up-regulated
1568787_at	Uncharacterized	Down-regulated
1563044_at	Uncharacterized	Down-regulated
1561144_at	Uncharacterized	Up-regulated

Despite limited information on the impact of these genes on IBD, conducting research that incorporates machine learning algorithms to analyze multiple datasets following the proposed methodology in this work holds promising results for researchers in this field. The findings presented in this work could potentially aid in the diagnosis and creation of new medical treatments, not only for IBD but also for other diseases.

The next steps are to compare REFS with advanced feature selection methods for biomarker discovery such as GRACES [48] and DNP [49]. Although AUC and MCC are considered standard metrics for bomarker discovery in the state of the art, additional metrics could be analyzed that could improve the results. Likewise, it is proposed to carry out external validations on different molecular data types and will include analysis of pathways from biomarker discovery to clinical application, validation including steps, regulatory considerations, and real-world applicability.

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] Q. Tang, X. Shi, Y. Xu, R. Zhou, S. Zhang, X. Wang, and J. Zhu, "Identification and Validation of the Diagnostic Markers for Inflammatory Bowel Disease by Bioinformatics Analysis and Machine Learning," Biochemical Genetics, 2023, pp. 1–14. doi.org/10.1007/s10528-023-10422-9.

[4] 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.

[5] W. Zhang, X. Chen, and K. C. Wong, "Noninvasive early diagnosis of intestinal diseases based on artificial intelligence in genomics and microbiome," Journal of Gastroenterology and Hepatology, 36(4), 2021, pp. 823–831. doi.org/10.1111/jgh.15500

[6] J. T. Rosenbaum, et al., "Revising the diagnosis of idiopathic uveitis by peripheral blood transcriptomics," American journal of ophthalmology, 222, 2021, pp. 15–23. doi.org/10.1016/j.ajo.2020.09.012

[7] H. Li, L. Lai, and J. Shen, "Development of a susceptibility gene based novel predictive model for the diagnosis of ulcerative colitis using random forest and artificial neural network," Aging, 12(20), 2020, pp. 20471–20482. doi.org/10.18632/aging.103861.

[8] T. B. Nguyen, D. N. Do, M. L. Nguyen-Thi, H. Hoang-The, T. T. Tran, and T. Nguyen-Thanh, "Identification of potential crucial genes and key pathways shared in Inflammatory Bowel Disease and cervical cancer by machine learning and integrated bioinformatics," Computers in Biology and Medicine, 149, 2022, p. 105996. doi.org/10.1016/j.compbiomed.2022.105996.

[9] T. He, et al., "Integrative computational approach identifies immune-relevant biomarkers in ulcerative colitis," FEBS Open bio, 12(2), 2022, pp. 500–515. doi.org/10.1002/2211-5463.13357.

[10] Y. Li, et al., "Screening of ulcerative colitis biomarkers and potential pathways based on weighted gene co-expression network, machine learning and ceRNA hypothesis," Hereditas 159(1), 2022, pp. 1–20. doi.org/10.1186/s41065-022-00259-4.

[11] L. Zhang, et al., "Identification of useful genes from multiple microarrays for ulcerative colitis diagnosis based on machine learning methods," Scientific reports, 12(1), 2022, p. 9962. doi.org/10.1038/s41598-022-14048-6.

[12] Z. A. Chen, H. H. Ma, Y. Wang, H. Tian, J. W. Mi, D. M. Yao, and C. J. Yang, "Integrated multiple microarray studies by robust rank aggregation to identify immune-associated biomarkers in Crohn's disease based on

three machine learning methods," Scientific Reports, 13(1), 2023, p. 2694. doi.org/10.1038/s41598-022-26345-1.

[13] H. M. Khorasani, H. Usefi, and L. Peña-Castillo, "Detecting ulcerative colitis from colon samples using efficient feature selection and machine learning," Scientific reports, 10(1), 2020, p. 13744. doi.org/10.1038/s41598-020-70583-0.

[14] X. Chen, et al., "Identification of FCN1 as a novel macrophage infiltration-associated biomarker for diagnosis of pediatric inflammatory bowel diseases," Journal of Translational Medicine 21(1), 2023, p. 203. doi.org/10.1186/s12967-023-04038-1.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] A. Lopez-Rincon et al., "Machine Learning-Based Ensemble Recursive Feature Selection of Circulating miRNAs for Cancer Tumor Classification," Cancers, 12(7), 2020, p. 1785. doi: 10.3390/cancers12071785.

[20] A. Lopez-Rincon, et al., "A robust mRNA signature obtained via Recursive Ensemble Feature Selection predicts the responsiveness of omalizumab in moderate-to-severe asthma," Authorea Preprints, 2022. DOI: 10.22541/au.166663264.45554320/v1.

[21] P. I. Metselaar, et al., "Recursive ensemble feature selection provides a robust mRNA expression signature for myalgic encephalomyelitis/chronic fatigue syndrome," Scientific reports 11(1), 2021, p. 4541. doi.org/10.1038/s41598-021-83660-9.

[22] S. Kidwai, et al., "USING MACHINE LEARNING FOR DRUG DISCOVERY IN IBD," CMBBE 2023-18th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering (CMBBE 2023), 2023. https://hal.science/hal-04097613.

[23] D. Rojas-Velazquez, A. Tonda, I. Rodriguez-Guerra, A. D. Kraneveld and A. Lopez-Rincon, "Multi-objective Evolutionary Discretization of Gene Expression Profiles: Application to COVID-19 Severity Prediction," In International Conference on the Applications of Evolutionary Computation (Part of EvoStar), 2023 pp. 703–717, Cham: Springer Nature Switzerland. doi.org/10.1007/978-3-031-30229-9_45.

[24] M. Benner, et al., "Antibiotic intervention affects maternal immunity during gestation in mice," Frontiers in Immunology, 12, 2021, p. 685742. doi.org/10.3389/fimmu.2021.685742.

[25] J. Blankestijn, et al., "Classifying asthma control using salivary and fecal microbiome in children with moderate to severe asthma: results from the SysPharmPediA study," European Respiratory Journal, 60(66), 2022. DOI: 10.1183/13993003.congress-2022.4554.

[26] D. Rojas-Velazquez, S. Kidwai, L. Cerezin, L. de Vries, K. Besermenji, P. Perez-Pardo, and A. Lopez-Rincon, "FEATURE SELECTION APPLIED TO MICROBIOME FOR DRUG DISCOVERY," In CMBBE 2023-18th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering (CMBBE 2023), 2023. https://hal.science/hal-04097608.

[27] D. Rojas-Velazquez, S. Kidwai, P. Perez-Pardo, D. Oberski, J. Garssen, A. D. Kraneveld, and A. Lopez-Rincon, "Towards a reduced feature selection pipeline in 16s rRNA microbiome data using Machine Learning," In The 9th Beneficial Microbes Conference, 2022. https://hal.science/hal-04097623.

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

[29] 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.

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

[31] J. Alsobrook, et al., "Novel genomic biomarkers that differentiate between inflammatory bowel disease and normal patients using peripheral blood specimens: 1125," Official journal of the American College of Gastroenterology— ACG, 103, 2008, p. S439.

[32] J. Alsobrook, L. Davis, T. Williams, J. Leighton, and C. Harris, " Accurate Differentiation Between Irritable Bowel Syndrome And Inflammatory Bowel Disease Using Novel Biomarkers from Peripheral Blood Specimens: 1322," Official journal of the American College of Gastroenterology— ACG, 104, 2009, p. S494.

[33] J. M. Moreira, G. Ohlsson, P. Gromov, R. Simon, G. Sauter, J. E. Celis, and I. Gromova, "Bladder cancer–associated protein, a potential prognostic biomarker in human bladder cancer," Molecular and Cellular Proteomics, 9(1), 2010, pp. 161–177. doi.org/10.1074/mcp.M900294-MCP200.

[34] F. Yuan, Y. H. Zhang, X. Y. Kong, and Y. D. Cai, "Identification of candidate genes related to inflammatory bowel disease using minimum redundancy maximum relevance, incremental feature selection, and the shortest-path approach," BioMed Research International, 2017. doi.org/10.1155/2017/5741948.

[35] M. Jarosz, M. Olbert, G. Wyszogrodzka, K. Młyniec, and T. Librowski, "Antioxidant and anti-inflammatory effects of zinc. Zinc–dependent NF-κB signaling," Inflammopharmacology, 25, 2017, pp. 11–24. doi.org/10.1007/s10787-017-0309-4.

[36] M. S. Rodriguez, et al., "The RING ubiquitin E3 RNF114 interacts with A20 and modulates NF–κB activity and T–cell activation," Cell death and disease 5(8), 2014, p. e1399. doi.org/10.1038/cddis.2014.366.

[37] L. Zhu, et al., "Role of RING-type E3 ubiquitin ligases in inflammatory signalling and inflammatory bowel disease," Mediators of Inflammation, 2020, pp. 1–10. doi.org/10.1155/2020/5310180.

[38] F. Palone, et al., "Fecal HMGB1 reveals microscopic inflammation in adult and pediatric patients with inflammatory bowel disease in clinical and endoscopic remission," Inflammatory bowel diseases, 22(12), 2016, pp. 2886–2893. doi.org/10.1097/MIB.0000000000000938.

[39] N. Das, et al., "HMGB1 activates proinflammatory signaling via TLR5 leading to allodynia," Cell reports, 17(4), 2016, pp. 1128–1140.

[40] R. Vitali, et al., "Fecal HMGB1 is a novel marker of intestinal mucosal inflammation in pediatric inflammatory bowel disease," Official journal of the American College of Gastroenterology— ACG, 106(11), 2011, pp. 2029–2040. DOI: 10.1038/ajg.2011.231.

[41] S. C. Sun, "A20 restricts inflammation via ubiquitin binding," Nature Immunology, 21(4), 2020, pp. 362–364. doi.org/10.1038/s41590-020-0632-6.

[42] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation", BMC genomics, 21(1), 2020, pp.1–13. doi.org/10.1186/s12864-019-6413-7.

[43] D. Rojas-Velazquez, et al., "Methodology for Biomarker Discovery with Reproducibility in Microbiome Data using Machine Learning," BMC Bioinformatics, 25(1), 2024, p. 26. https://doi.org/10.1186/s12859-024-05639-3

[44] I. Arijs, et al., "Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment," PloS one, 4(11), 2009, e7984.

[45] I. Arijs, et al., "Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis," Gut, 58(12), 2009, pp. 1612–1619.

[46] A. Lopez-Rincon, et al., "Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning," Scientific reports, 11(1), 2021, p. 947. https://doi.org/10.1038/s41598-020-80363-5

[47] C. A. Perez-Romero, et al., "An Innovative AI-based primer design tool for precise and accurate detection of SARS-CoV-2 variants of concern," Sci Rep 13, 2023, p. 15782. https://doi.org/10.1038/s41598-023-42348-y

[48] C. Chen, S. T. Weiss, Y.-Y. Liu, "Graph convolutional network-based feature selection for high-dimensional and low-sample size data," Bioinformatics 39 (4) (2023) btad135. https://doi.org/10.1093/bioinformatics/btad135

[49] B. Liu, Y. Wei, Y. Zhang, Q. Yang, "Deep neural networks for high dimension, low sample size data," in: IJCAI, 2017, pp. 2287–2293https://doi.org/10.1093/bioinformatics/btad135