Multi-Objective Evolutionary Discretization of Gene Expression Profiles: Application to COVID-19 Severity Prediction

David Rojas-Velazquez, Alberto Tonda, Itzel Rodriguez-Guerra, Aletta Kraneveld, Alejandro Lopez-Rincon

Multi-Objective Evolutionary Discretization of Gene Expression Profiles: Application to COVID-19 Severity Prediction

David Rojas-Velazquez[1,2][0000−0003−4402−4736], Alberto Tonda[3,4][0000−0001−5895−4809], Itzel Rodriguez-Guerra[5], Aletta D. Kraneveld[1][0000−0001−9819−383X], and Alejandro Lopez-Rincon[1,2][0000−0003−4491−5889]

[1] Division of Pharmacology, University of Utrecht, Universiteitsweg 99, 3584 CG, Utrecht, The Netherlands
{e.d.rojasvelazquez,A.D.Kraneveld,a.lopezrincon}@uu.nl
[2] Department of Data Science, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
[3] UMR 518 MIA-PS, INRAE, Université Paris-Saclay, 91120, Palaiseau, France
alberto.tonda@inrae.fr
[4] Institut des Systèmes Complexes de Paris Île-de-France (ISC-PIF) - UAR 3611 CNRS, Paris, France
[5] Centro de Investigaciones en Ciencias Microbiologícas, Instituto de Ciencias, Benemérita Universidad Autónoma de Puebla IC 11 Puebla, Puebla, Mexico
Guadalupe.rodriguezgue@alumno.buap.mx

Abstract. Machine learning models can use information from gene expressions in patients to efficiently predict the severity of symptoms for several diseases. Medical experts, however, still need to understand the reasoning behind the predictions before trusting them. In their day-to-day practice, physicians prefer using *gene expression profiles*, consisting of a discretized subset of all data from gene expressions: in these profiles, genes are typically reported as either over-expressed or under-expressed, using discretization thresholds computed on data from a healthy control group. A discretized profile allows medical experts to quickly categorize patients at a glance. Building on previous works related to the automatic discretization of patient profiles, we present a novel approach that frames the problem as a multi-objective optimization task: on the one hand, after discretization, the medical expert would prefer to have as few different profiles as possible, to be able to classify patients in an intuitive way; on the other hand, the loss of information has to be minimized. Loss of information can be estimated using the performance of a classifier trained on the discretized gene expression levels. We apply one common state-of-the-art evolutionary multi-objective algorithm, NSGA-II, to the discretization of a dataset of COVID-19 patients that developed either mild or severe symptoms. The results show not only that the solutions found by the approach dominate traditional discretization based on statistical analysis and are more generally valid than those obtained through single-objective optimization, but that the candidate Pareto-optimal so-

lutions preserve the sense-making that practitioners find necessary to trust the results.

Keywords: Gene Expressions · Patient Profiles · Multi-Objective Evolutionary Algorithms · COVID-19.

1 Introduction

The information available up to November 16, 2022 indicates that the SARS-CoV-2 pandemic continues, with 640 million cases and over 6 million deaths [37]. Due to the magnitude of the viral outbreak, one of the great problems that humanity faces is the lack of medical equipment to care for infected patients. Several studies have tried to predict the difference between the severity of the cases using machine learning to analyze datasets with Chest X-Ray images, but each dataset needs to be validated by experts, annotated with the corresponding lesions of lung diseases, and features extracted based on recommendations of medical personnel [1] resulting in the information taking time to be available for analysis. One alternative to avoid the problem with Chest X-Ray images is to use omics data, for example using DNA methylation [31, 24], mRNA gene expression [16, 32] and/or microRNA [44] data, to quickly anticipate if a patient will be in need of intensive care, and efficiently distribute the available medical resources. Several sources [7, 39] consider the correct management of beds and the resources available in hospitals to be a crucial factor in reducing mortality rates from COVID-19 in patients with severe infections.

Although there have been varying degrees of success with the use of multi-omic data for diagnostic and prognostic purposes in general, one of the challenges lies in translating the results into meaningful diagnostic tests or biomarkers for clinical practice [11]. Nowadays multiple mRNA gene expression datasets are available in public repositories: typically gene expression data will include thousands of genes (features) related to just a few samples, leading to several challenges for finding meaningful biomarkers. As humans cannot process the information contained in thousands of genes due to the complexity of the data, it is necessary to use computational tools such as machine learning (ML) techniques to obtain reliable predictions [26, 27, 36].

mRNA gene expression datasets are matrices generated by sequencing data, where typically each column corresponds to a specific variant of a gene, and each row to a sample from a tissue (e.g. peripheral blood mononuclear cells, peripheral blood leukocytes, whole blood, etc.). The values in the matrix are the expression levels of the given variant of a gene for a patient sample [3].

Nevertheless, to make sense of the data, medical practitioners often resort to the creation of *gene expression profiles*, discretizations of gene expressions, where each continuous value is assigned to a discrete category, for example under- or over-expressed gene. In this work, patient profile is defined as a set of gene expression values that uniquely characterize the patient, often discretized to make it more readable by domain experts. Two patients have the same patient profile

if their discretized values are the same for the expression of each considered gene. Categories are usually evaluated using thresholds based on the mean values of gene expressions from healthy controls as a baseline. While this procedure can help the sensemaking of the experts, such a discretization leads to loss of information and could potentially impair classification performance. Furthermore, relying on control groups for discretizing gene expressions can lead to the wrong conclusions, as gene expression variability can be high, and control groups are usually comparatively small.

In this paper, we frame the problem of discretizing gene expression profiles from patients as a multi-objective task: on the one hand, the aim is to deliver a discretization that can be easily interpreted by a domain expert, ideally minimizing the different types of patient profiles; on the other hand, the loss of information resulting from the discretization should be minimal. While counting the number of different patient profiles resulting from the discretization is trivial, the loss of information can be assessed through the performance of a classifier in cross-validation on the discretized patient profiles. While approaches for the automated discretization of patient profiles have already been proposed in literature [30], the problem was previously conceived as single-objective optimization, with an arbitrary choice of weights to find compromises between conflicting objectives.

To test the proposed approach, real data from 138 participants, including information from 60,671 genes, were used and compared with a classical discretization approach based on mean values of gene expression from the group of healthy controls, and the single best solution found by a previously presented single-objective automated approach. Experimental results show that the proposed methodology is effective, identifying 12 genes highly correlated with the response to treatment and being able to discretize their gene expression levels into gene expression profiles. This helps to increase the performance of classifiers, and at the same time provides a human-interpretable explanation of the development of mild or dire symptoms from a COVID-19 infection. An expert analysis performed by domain experts provides a final validation of our approach.

2 Background

This section provides the minimal information needed to introduce the scope of our work.

2.1 Feature Selection

In machine learning (ML), feature selection (FS) is defined as the process of choosing the features of a dataset in order to obtain a minimal, informative subset. Features may not be part of this subset for two main reasons: they might be unrelated to the underlying nature of the problem, just adding noise; or they might be heavily correlated with other features, adding no relevant information for the task. Applications range from face recognition [43] to medicine [49],

and approaches can be divided into two categories [19]: filters that score features according to a criterion (often a statistical test); and recursive procedures (forward or backwards) that attempt to reduce the features to a small set of non-redundant ones [25, 10].

In the scope of this work, we focus on recursive FS algorithms, in particular Recursive Ensemble Feature Selection (REFS). The method is a variation of Recursive Feature Elimination (RFE) [20], scoring the features in a 10-fold cross-validation scheme, using 8 different classifiers: Gradient Boosting [18], Passive Aggressive Classifier [13], Logistic Regression [12], Support Vector Machines Classifier [35], Random Forest [5], Stochastic Gradient Descent on Linear Models [48], Ridge Classifier [21] and Bagging [4]. The 10-fold cross-validation scheme was implemented following the nested cross-validation approach described in [42], which proves to be an effective approach to avoid the overfitting of machine learning models when working with data sets with a small number of samples [42]. The lowest-scoring features are removed from the analysis and the process is repeated until the overall classification accuracy drops below a given threshold. The use of an ensemble of classifiers reduces the effects of the inherent bias in each ML algorithm, thus delivering a more objective feature ranking. This technique has been applied successfully for problems involving both mRNA [27] and miRNA [26], featuring number of features ranging from 1,046 to 54,675.

2.2 Gene Expression Profiles

For diagnostic purposes it is not uncommon to generate heatmaps via computational techniques such as clustering or univariate analysis to find genetic expression profiles [14, 28, 33], using healthy controls as a baseline to obtain discretization thresholds. A visual representation of the transformation from a gene expression dataset to gene expression profiles is shown in Fig. 1.

Nevertheless, the generated genetic profiles or found biomarkers are difficult to translate into clinical practice [11]. One of the reasons is that not every time it is possible to include healthy controls as a baseline in the studies, and therefore reference values from other studies should be used which could be affected by the batch effect [36]. Furthermore, control groups are usually small in size, while the variability of gene expression can be considerable. Automated discretization could be a viable alternative, but approaches in literature [30] only frame the problem as single-objective, while in reality there are two conflicting objectives: maximizing classification performance and minimizing the number of different patient profiles to be analyzed by a human expert.

3 Proposed approach

In this section, a novel approach for the multi-objective discretization of gene expression data to obtain gene expression profiles, interpretable by domain experts is presented. After performing a step of feature selection to identify the

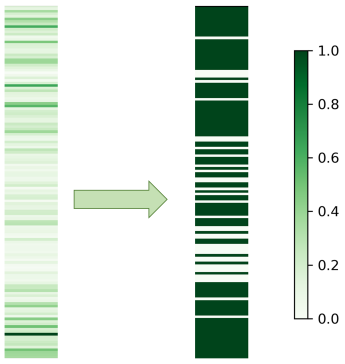

Fig. 1. Example of discretization of gene expression data to gene expression profiles for one gene, using a threshold value to distinguish between over-expressed (dark green) and under-expressed (light green) genes.

most relevant genes, these identified genes are discretized using thresholds optimized by a multi-objective evolutionary algorithm (MOEA), The conflicting objectives to be optimized are the classification performance and the number of different profiles (rows with different values) in the discretized dataset.

3.1 Feature Selection

For the feature selection step, the REFS algorithm was executed. REFS is an algorithm that uses the feedback of an ensemble of classifiers to rank each feature depending on its use capacity for the process of classification. The lowest-scoring features are removed, and the classification/ranking is repeated, until the average classification accuracy is below to a threshold defined by the user. The objective of this process is to select the most meaningful genes to correctly predict and model COVID-19 patients' severity (mild/severe).

3.2 Multi-Objective Evolutionary Discretization

After running the REFS algorithm, a small set of relevant features is obtained. In this case, the values associated to the relevant features are complex and difficult to read for medical decision making. So, instead of showing each feature as a continuous value, we use a MOEA to categorize the values into underexpression and overexpression, optimizing the thresholds for each selected feature (gene). Given the reduced set of features $F = \{f_0, f_1, f_2, ..., f_n\}$, given by the REFS algorithm, EAs was used to transform input variables into *over* and *under* expressed values, labeled as 0 and 1, respectively: that is, EA will generate a vector of thresholds $I = \{t_0, t_1, t_2, ..., t_n\}$ to discretize each variable. Two criteria were used to optimize the discretization: number of different profiles, to be minimized

to support the sensemaking of domain experts; and classification performance (given the labels of the dataset, corresponding to mild or severe symptoms), to be maximized, as a proxy of information loss. Consequently, the fitness functions for an individual I are given by:

$$f_1(I) = \frac{1.0}{1.0 + F1_{cv}(X_i, y)} \qquad f_2(I) = n_p \tag{1}$$

where X_i is the dataset discretized according to the thresholds of individual I, y is the vector with the labels (mild/severe symptoms) for each sample, $F1_{cv}(X, y)$ is the F1 score in a cross-validation, n_p is the number of different profiles in the dataset after discretization. Previous works on automated discretization [30] employed accuracy as part of the evaluation, but F1 (a number between 0 and 1, representing the harmonic mean between precision and recall) is a preferable metric in case of unbalanced datasets, where samples from one class are more prevalent, as is often the case for medical data. The fitness function should be minimized, since the ideal candidate solution presents a high F1 and a low number of different profiles, to facilitate the understanding of domain experts.

4 Experimental evaluation

The proposed approach is implemented in Python 3, relying on the `inspyred`[6] package for NSGA-II and the `scikit-learn` [34] package for classification. All the code and data needed to reproduce the experiments is freely available in the GitHub repository: https://github.com/to-be-disclosed/after-review.

4.1 Data

The dataset GSE169687 was selected from the gene expression omnibus (GEO) repository [7]. This dataset contains 138 samples of mRNA from peripheral blood of recovered COVID-19 patients at different time points, and 14 healthy controls. Only the 138 samples from patients with either mild/moderate (n=109) or severe/critical (n=29) symptoms were considered in the experiments, while the information from the 14 healthy controls was later used to compute an expert candidate solution to compare against. There are 60,671 ensemble genes (features) for each sample, and the dataset was divided into 2 classes, where label 0 was assigned to patients with mild/moderate symptoms and label 1 to patients with severe/critical symptoms.

[6] https://pythonhosted.org/inspyred/
[7] https://www.ncbi.nlm.nih.gov/geo/

4.2 Feature Selection

REFS algorithm was run 10 times, and a reduction from 60,671 to 12 features (highlighted as the optimal trade-off in Fig. 2) was ultimately obtained. This translates to the expression levels shown in Fig. 3.

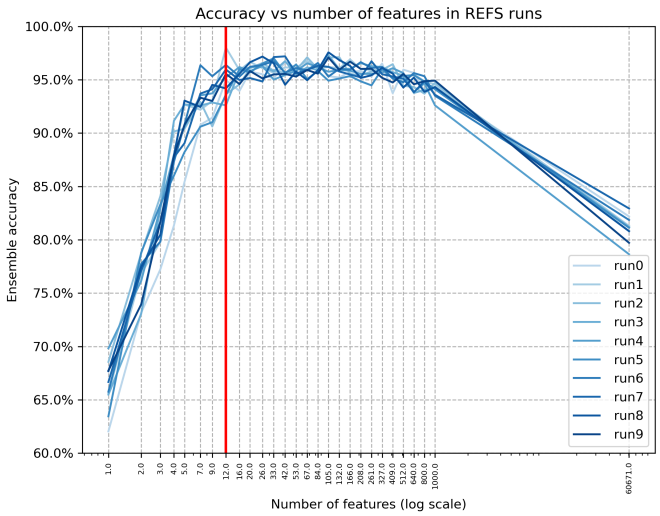

Fig. 2. 10 runs of the REFS algorithm. The solution considered as the best compromise between accuracy and number of features is marked with a red line (n=12).

4.3 Profile Generator

The MOEA selected for the profile generator is the NSGA-II [15], considered among the state of the art for multi-objective optimization with a relatively low number of objectives. After a few trial runs, the algorithm is set with the following parameters: $\mu = 200$, $\lambda = 350$, $p_c = 0.8$, $p_m = 0.2$, stop condition after 100 generations, and Logistic Regression [12] as the classifier chosen to compute classification performance $F1_{cv}$ for the fitness function described in Eq. 1. The choice of Logistic Regression is motivated by its effectiveness and training speed, making it one of the most suitable algorithms for our scenario. The classifier is run in a 10-fold cross-validation at each evaluation, in order to obtain a more reliable estimate of F1. The whole evolutionary optimization process is repeated 30 times, to assess the variance in the final results.

To provide a comparison, profiles based on a classical technique of the domain were compared, using as a reference the gene expression levels of the healthy controls to discretize the gene expressions of the patients: in other words, if a

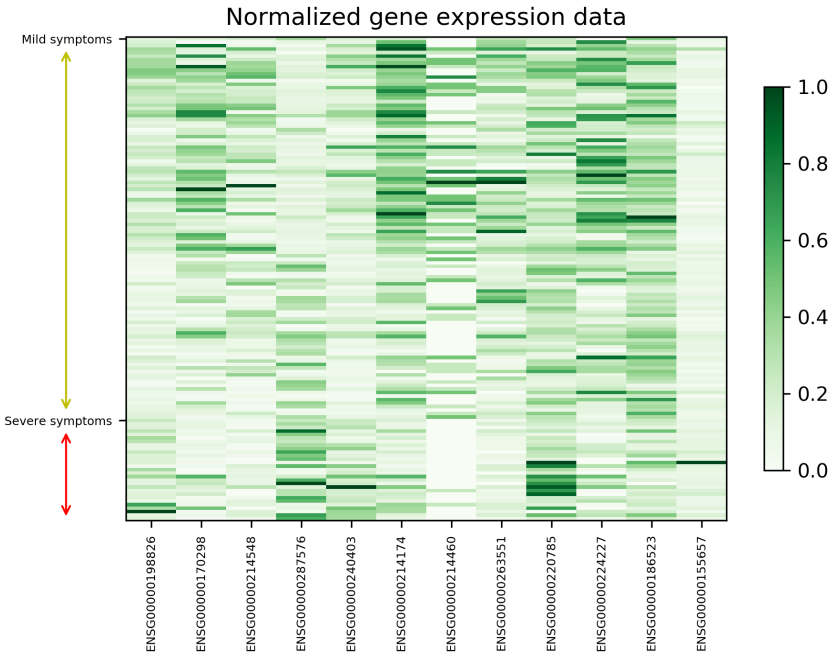

Fig. 3. Heatmap of normalized gene expression data, showing the values of each patient for the 12 most important genes (labeled as ENSG*) selected by the REFS algorithm.

patient has the gene expression for a given gene higher than the mean of the 14 healthy controls in the dataset, that gene expression will be categorized as over-expressed (label 1); otherwise, it will be categorized as under-expressed (label 0).

The results of the experiments are shown in Figure 4: it is clear that the point corresponding to the classical discretization technique (orange triangle) is dominated by the candidate solutions found by the proposed approach. On the other hand, it is interesting to observe that the point corresponding to the single-objective automated discretization [30] (red cross) is Pareto-optimal, and covers a part of the objective space not explored by the MOEA. In order to properly compare the proposed approach to other methods, two candidate solutions on the Pareto front are selected: the one with the best value of $F1_{cv}$ ($F1_{cv} = 0.9457$, $n_p = 40$); and the one with the fewest profiles that still scored above an arbitrary threshold $F1_{cv} \geq 0.9$ ($F1_{cv} = 0.9257$, $n_p = 15$), considered acceptable by an expert. They will be used in all comparisons in the following.

A possible disadvantage of the automatic methodologies is the generation of discretization thresholds that fit only to the classifier used for the fitness function (Logistic Regression) and lose generality. To test the generality of the method, the data was transformed using all the resulting thresholds from the two MOEA-selected solutions, the expert solution, and the best solution found by the single-objective approach. The mean accuracy in a 10-fold cross-validation was computed using Logistic Regression and seven other state-of-the-art classifiers: Passive Aggressive Classifier [13], Stochastic Gradient Descent on Linear Models [48], Support Vector Machines [35], Gradient Boosting [18], Random Forest [5], Ridge Classifier [21] and Bagging [4]. The results presented in Table 1 show that the discretizations found through automated approaches all perform better than the expert solution. Furthermore, both solutions obtained through the proposed MOEA approach on average perform better than the one found through single-objective optimization, hinting that the latter might be overfitted to the performance of Logistic Regression.

Table 1. Discretization strategies compared, using the F1 from different state-of-the-art classifiers, in a 10-fold cross-validation. **MOEA Highest F1** is the candidate solution with the highest F1 on the Pareto front. **MOEA Fewest Profiles** is the Pareto-optimal solution with the fewest profiles. **Best single-objective** is the performance of the best individual produced by the single-objective optimizer in [30]. **Expert solution** indicates the discretization performed using healthy controls as reference for the thresholds in the gene expression levels.

Classifier	**MOEA Highest F1 (40 profiles)**	**MOEA Fewest profiles (15 profiles)**	**Best single-objective (46 profiles)**	**Expert solution (122 profiles)**
BaggingClassifier	0.8990 +/- 0.1344	0.9114 +/- 0.0908	0.8883 +/- 0.1414	0.6633 +/- 0.2406
GradientBoostingClassifier	0.9314 +/- 0.0859	0.9257 +/- 0.0923	0.8362 +/- 0.1543	0.6190 +/- 0.2376
LogisticRegression	0.9457 +/- 0.0842	0.9257 +/- 0.0923	0.9800 +/- 0.0600	0.7667 +/- 0.1633
PassiveAggressiveClassifier	0.8324 +/- 0.1976	0.7267 +/- 0.3072	0.8455 +/- 0.1213	0.6943 +/- 0.2615
RandomForestClassifier	0.9314 +/- 0.0859	0.9457 +/- 0.0842	0.8924 +/- 0.1165	0.6824 +/- 0.2475
RidgeClassifier	0.8467 +/- 0.3027	0.8657 +/- 0.2967	0.9000 +/- 0.1000	0.7667 +/- 0.1633
SVC	0.9514 +/- 0.0756	0.9257 +/- 0.0923	0.9457 +/- 0.0842	0.8029 +/- 0.0977
Mean F1	0.9252	0.9059	0.8892	0.6990

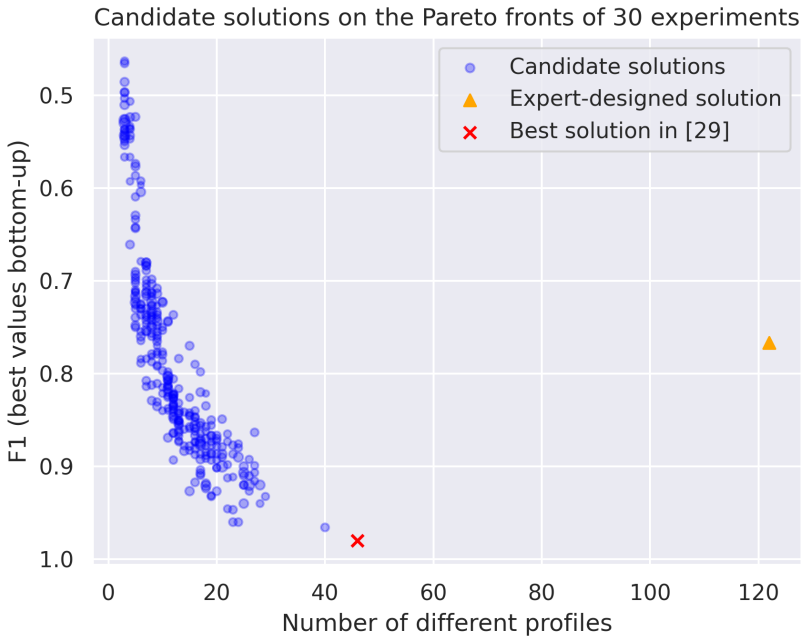

Fig. 4. Experimental results, with number of profiles on the x-axis, best values towards the left, and F1 on the y-axis, best values towards the bottom. Points in blue represent candidate solutions found during the runs. The orange triangle corresponds to the expert discretization based on gene expressions of healthy controls ($F1_{cv} = 0.77$, $n_p = 122$). The red cross corresponds to the best value found by the single-objective method presented in [30] ($F1_{cv} = 0.98$, $n_p = 46$).

Fig. 5 shows the heatmaps obtained by discretizing the original dataset using the two solutions selected on the Pareto front: note how some of the discretized features are homogeneous, with all or nearly all values assigned to the same class. Analyzing the values of the thresholds reported in Table 2, it is interesting to notice how the MOEA, in order to reduce the number of profiles, set some of the thresholds to 1.0, *de facto* assigning almost all values of that feature (values between 0.0 and 1.0) to the same class. From a certain point of view, it's as if the algorithm were performing a second feature selection. An inspection of all the Pareto-optimal solutions found during the 30 experiments found that the most common thresholds set to 1.0 are for ENSG00000198826 (gene ARHGAP11A), ENSG00000214174 (gene AMZ2P1), and ENSG00000186523 (gene FAM86B1).

Table 2. Generated thresholds for each of the genes.

Ensemble ID	Gene ID	Thresholds (highest F1)	Thresholds (fewest profiles)	Thresholds (single objective)
ENSG00000198826	ARHGAP11A	0.0342	1.0000	0.2869
ENSG00000170298	LGALS9B	0.1654	0.5804	0.2817
ENSG00000214548	MEG3	0.1592	1.0000	0.6731
ENSG00000287576	-	0.5799	0.5858	0.4580
ENSG00000240403	KIR3DL2	0.0846	0.2320	0.1622
ENSG00000214174	AMZ2P1	1.0000	1.0000	0.2013
ENSG00000214460	TPT1P6	0.4441	1.0000	0.1101
ENSG00000263551	-	0.2366	0.1977	0.2379
ENSG00000220785	MTMR9LP	0.6698	0.9265	0.6528
ENSG00000224227	OR2L1P	1.0000	0.0267	0.9902
ENSG00000186523	FAM86B1	1.0000	1.0000	0.9191
ENSG00000155657	TTN	0.6623	0.7013	0.9880

5 Discussion

The results described in Table 1 show that the proposed approach outperforms the expert-driven discretization methodology from both the classification performance, and the number of profiles produced after discretization. Additionally, the high accuracy results obtained in a 10-fold cross-validation for several classifier provide evidence that using only Logistic Regression as element of the fitness function does not produce overfit in the discretization thresholds to a single classifier, contrary to the single-objective discretization.

Two of the 12 genes selected by the proposed approach are novel transcripts: ENSG00000263551 and ENSG00000287576. These two genes are listed as lncRNA (long, non-coding RNA) in the gene cards database [38], and there is no information available related to the subject as well as with ENSG00000214460 (TPT1P6 gene). These findings could be an important lead for new research on the subject, as they have never been associated with any particular biological function in literature, to our knowledge.

AMZ2P1, KIR3DL2and LGALS9B genes are directly related to the severity of symptoms in COVID. AMZ2P1 was found to be over-expressed in healthy

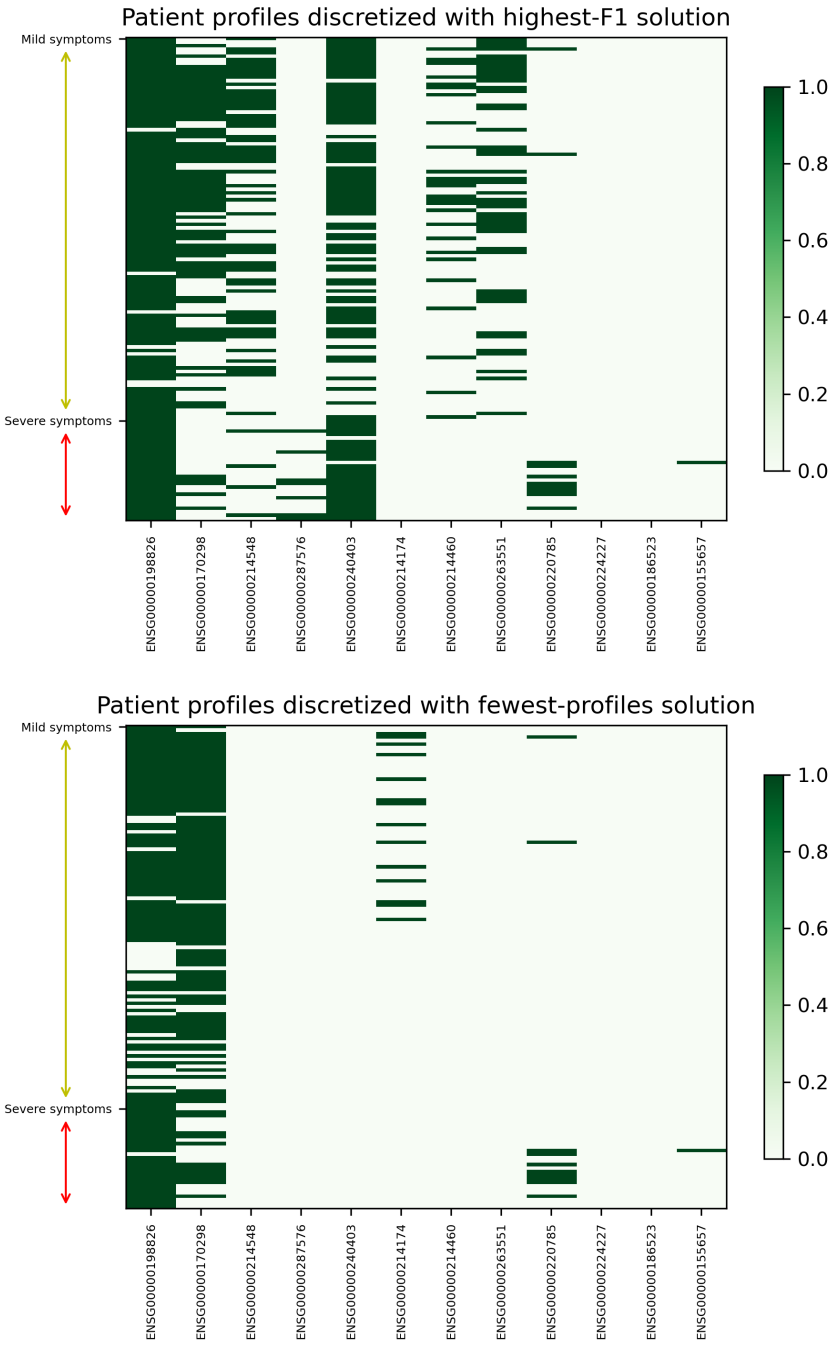

Fig. 5. Heatmaps produced by applying the set of thresholds identified by the two selected Pareto-optimal solutions: highest F1 (top) and fewest profiles (bottom).

retesting-positive COVID-19 patients [17]. KIR3DL is a killer cell immunoglobulin-like receptor gene and it's over-expression in presence of HLA-B is associated with moderate COVID-19 [2]. LGALS9 was identified as a COVID-19 severity protein biomarker, further evaluation of this gene provided evidence that it is also implicated in the apoptosis-associated cytokine Fas cell surface death receptor as a causal mediator of severe COVID-19 [22]. Although, not related to COVID-19 severity, we found two genes that are directly related to COVID-19 symptoms: TTN and OR2L1P. TTN encodes a large abundant protein of striated muscle. This gene is related to the molecular mechanisms behind muscle loss in COVID and this is associated with altered regulation of several cytokines [6], this could explain the muscle fatigue present in COVID-19 patients. OR2LP1 encodes an olfactory receptor, which interacts with odorant molecules in the nose, connected to the inflammatory reaction in the nasal cavity due to COVID, that leads to a temporary anosmia, where the odors are not able to reach the olfactory receptor neurons [40]; loss of the sense of smell is a common COVID-19 symptom [23]

Genes MEG3, FAM86B1, MTMR9LP and ARHGAP11A do not appear as biomarkers directly related to COVID-19. However, the over-expression of MEG3 may be favorable to virus infection, as evidenced in influenza A, via ADAR over-expression [8]. Since ADAR has been found as a controlling element in cellular response to viral infections, its regulation by MEG3 and interactions with lncRNAs in SARS-CoV-2 infected cells may influence progression of the disease [41]. MTMR9LP is a lncRNA that is under-expressed in cryptococcal meningitis patients in comparison of healthy controls, related with cytokine expression and immune response triggered by cryptococcal infection [47].

ARHGAP11A is a member of the Rho GTPase-activating protein (RhoGAP) subfamily: this gene is under-expressed in tumors and is usually associated with malignant progression, this may be due the ability of ARHGAP11 to physically bind to p53 and promote its function, to induce cell-cycle arrest and apoptosis [45]. Interestingly, it has been found that a highly expressed ARHGAP11 is a sign for bad prognosis and poor survival rate in lung adenocarcinoma [9]. Also, this gene is under-expressed in pulmonary arterial hypertension, and this could be related to its role as a regulator of cell cycle-dependent motility [29]. FAM86B1 is a gene proposed as a "dark gene" directly linked with the survival of patients in complex diseases, specifically in bladder urothelial carcinoma (BLCA) [46]. Nevertheless, ARHGAP11A, AMZ2P1 and FAM86B1 probably provide less information, as inferred from the thresholds set by the MOEA, that are often set to 1.0 for these genes among the Pareto-optimal solutions.

6 Conclusions and Future Works

In this paper, a novel multi-objective evolutionary approach to the discretization of gene expression data was presented to obtain interpretable gene expression profiles, this approach can also lead to good classification accuracy when used with ML classifiers. The results on a real-world dataset related to COVID-19

(with patients exhibiting either mild or severe symptoms) seem promising, showing that the proposed technique performs better than a more classical approach based on a comparison with healthy controls, and produces results with better generality than a previous single-objective approach. In addition, we generated a set of rules given 12 specific genes to be used as a guide to decide whether a patient will present severe symptoms. An expert analysis reveals that the genes identified by our approach are known in literature, and the experts are satisfied by the discretization options provided, with a preference for solutions producing fewer profiles. While the initial results are promising, tests are needed in other real-world databases related to COVID-19 patients. Additionally, in order to claim generality, the proposed approach still needs to be evaluated on datasets related to different diseases.

References

1. Alghamdi, H.S., Amoudi, G., Elhag, S., Saeedi, K., Nasser, J.: Deep learning approaches for detecting covid-19 from chest x-ray images: A survey. Ieee Access **9**, 20235–20254 (2021)
2. Bernal, E., Gimeno, L., Alcaraz, M.J., Quadeer, A.A., Moreno, M., Martínez-Sánchez, M.V., Campillo, J.A., Gomez, J.M., Pelaez, A., García, E., et al.: Activating killer-cell immunoglobulin-like receptors are associated with the severity of covid-19. The Journal of Infectious Diseases (2021)
3. Brazma, A., Vilo, J.: Gene expression data analysis. FEBS letters **480**(1), 17–24 (2000)
4. Breiman, L.: Pasting small votes for classification in large databases and on-line. Machine Learning **36**(1-2), 85–103 (1999)
5. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
6. Cantu, N., Vyavahare, S., Kumar, S., Chen, J., Kolhe, R., Isales, C.M., Hamrick, M., Fulzele, S.: Synergistic effects of multiple factors involved in covid-19-dependent muscle loss. Aging and Disease pp. 9–9 (2021)
7. Cavallo, J.J., Donoho, D.A., Forman, H.P.: Hospital capacity and operations in the coronavirus disease 2019 (COVID-19) pandemic—planning for the nth patient. JAMA Health Forum **1**(3), e200345 (Mar 2020). https://doi.org/10.1001/jamahealthforum.2020.0345, https://doi.org/10.1001/jamahealthforum.2020.0345
8. de Chassey, B., Aublin-Gex, A., Ruggieri, A., Meyniel-Schicklin, L., Pradezynski, F., Davoust, N., Chantier, T., Tafforeau, L., Mangeot, P.E., Ciancia, C., et al.: The interactomes of influenza virus ns1 and ns2 proteins identify new host factors and provide insights for adar1 playing a supportive role in virus replication. PLoS pathogens **9**(7), e1003440 (2013)
9. Chen, S., Duan, H., Xie, Y., Li, X., Zhao, Y.: Expression and prognostic analysis of rho gtpase-activating protein 11a in lung adenocarcinoma. Annals of Translational Medicine **9**(10) (2021)
10. Chien, Y., Fu, K.S.: On the generalized karhunen-loève expansion (corresp.). IEEE Transactions on Information Theory **13**(3), 518–520 (1967)
11. Chin, L., Gray, J.W.: Translating insights from the cancer genome into clinical practice. Nature **452**(7187), 553–563 (2008)
12. Cox, D.R.: The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological) pp. 215–242 (1958)

13. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online passive-aggressive algorithms. Journal of Machine Learning Research **7**(Mar), 551–585 (2006)
14. Cruz-Rodriguez, N., Quijano, S.M., Enciso, L.J., Combita, A.L., Zabaleta, J.: Gene expression signature predicts induction treatment response and clinical outcome in adult colombian patients with acute lymphoblastic leukemia (2016)
15. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE transactions on evolutionary computation **6**(2), 182–197 (2002)
16. Delorey, T.M., Ziegler, C.G., Heimberg, G., Normand, R., Yang, Y., Segerstolpe, Å., Abbondanza, D., Fleming, S.J., Subramanian, A., Montoro, D.T., et al.: Covid-19 tissue atlases reveal sars-cov-2 pathology and cellular targets. Nature pp. 1–8 (2021)
17. Fang, K.Y., Liang, G.N., Zhuang, Z.Q., Fang, Y.X., Dong, Y.Q., Liang, C.J., Chen, X.Y., Guo, X.G.: Screening the hub genes and analyzing the mechanisms in discharged covid-19 patients retesting positive through bioinformatics analysis. Journal of Clinical Laboratory Analysis **36**(7), e24495 (2022)
18. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
19. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (Mar 2003)
20. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning **46**(1-3), 389–422 (2002)
21. Hoerl, A.E., Kennard, R.W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics **12**(1), 55–67 (1970). https://doi.org/10.1080/00401706.1970.10488634
22. Klaric, L., Gisby, J.S., Papadaki, A., Muckian, M.D., Macdonald-Dunlop, E., Zhao, J.H., Tokolyi, A., Persyn, E., Pairo-Castineira, E., Morris, A.P., et al.: Mendelian randomisation identifies alternative splicing of the fas death receptor as a mediator of severe covid-19. medRxiv (2021)
23. Klopfenstein, T., Kadiane-Oussou, N., Toko, L., Royer, P.Y., Lepiller, Q., Gendrin, V., Zayet, S.: Features of anosmia in covid-19. Medecine et maladies infectieuses **50**(5), 436–439 (2020)
24. Konigsberg, I.R., Barnes, B., Campbell, M., Davidson, E., Zhen, Y., Pallisard, O., Boorgula, M.P., Cox, C., Nandy, D., Seal, S., et al.: Host methylation predicts sars-cov-2 infection and clinical outcome. Communications medicine **1**(1), 1–10 (2021)
25. Lewis, P.: The characteristic selection problem in recognition systems. IRE Transactions on information theory **8**(2), 171–178 (1962)
26. Lopez-Rincon, A., Martinez-Archundia, M., Martinez-Ruiz, G.U., Schoenhuth, A., Tonda, A.: Automatic discovery of 100-mirna signature for cancer classification using ensemble feature selection. BMC bioinformatics **20**(1), 480 (2019)
27. Lopez-Rincon, A., Mendoza-Maldonado, L., Martinez-Archundia, M., Schönhuth, A., Kraneveld, A.D., Garssen, J., Tonda, A.: Machine learning-based ensemble recursive feature selection of circulating mirnas for cancer tumor classification. Cancers **12**(7), 1785 (2020)
28. Lu, Y., Fang, Z., Li, M., Chen, Q., Zeng, T., Lu, L., Chen, Q., Zhang, H., Zhou, Q., Sun, Y., et al.: Dynamic edge-based biomarker non-invasively predicts hepatocellular carcinoma with hepatitis b virus infection for individual patients based on blood testing. Journal of molecular cell biology **11**(8), 665–677 (2019)

29. Ma, Y., Chen, S.S., Feng, Y.Y., Wang, H.L.: Identification of novel biomarkers involved in pulmonary arterial hypertension based on multiple-microarray analysis. Bioscience Reports **40**(9) (2020)
30. Mouhrim, N., Tonda, A., Rodríguez-Guerra, I., Kraneveld, A.D., Rincon, A.L.: An evolutionary approach to the discretization of gene expression profiles to predict the severity of COVID-19. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion. ACM (Jul 2022). https://doi.org/10.1145/3520304.3529001, https://doi.org/10.1145/3520304.3529001
31. de Moura, M.C., Davalos, V., Planas-Serra, L., Alvarez-Errico, D., Arribas, C., Ruiz, M., Aguilera-Albesa, S., Troya, J., Valencia-Ramos, J., Vélez-Santamaria, V., et al.: Epigenome-wide association study of covid-19 severity with respiratory failure. EBioMedicine **66**, 103339 (2021)
32. Ng, D.L., Granados, A.C., Santos, Y.A., Servellita, V., Goldgof, G.M., Meydan, C., Sotomayor-Gonzalez, A., Levine, A.G., Balcerek, J., Han, L.M., et al.: A diagnostic host response biosignature for covid-19 from rna profiling of nasal swabs and blood. Science advances **7**(6), eabe5984 (2021)
33. Paiva, B., Corchete, L.A., Vidriales, M.B., Puig, N., Maiso, P., Rodriguez, I., Alignani, D., Burgos, L., Sanchez, M.L., Barcena, P., et al.: Phenotypic and genomic analysis of multiple myeloma minimal residual disease tumor cells: a new model to understand chemoresistance. Blood, The Journal of the American Society of Hematology **127**(15), 1896–1906 (2016)
34. 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. the Journal of machine Learning research **12**, 2825–2830 (2011)
35. Platt, J., Others: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
36. Rincon, A.L., Kraneveld, A.D., Tonda, A.: Batch correction of genomic data in chronic fatigue syndrome using CMA-ES. In: Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion. pp. 277–278. ACM, Cancun, Mexico (Jul 2020). https://doi.org/10.1145/3377929.3389947, https://doi.org/10.1145/3377929.3389947
37. Roser, M.: Covid-19 data explorer (2022), https://ourworldindata.org/explorers/coronavirus-data-explorer
38. Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., et al.: Genecards version 3: the human gene integrator. Database **2010** (2010)
39. Sussman, N.: Time for bed(s): Hospital capacity and mortality from covid-19. COVIDEconomics pp. 116–129 (2020)
40. Torabi, A., Mohammadbagheri, E., Akbari Dilmaghani, N., Bayat, A.H., Fathi, M., Vakili, K., Alizadeh, R., Rezaeimirghaed, O., Hajiesmaeili, M., Ramezani, M., et al.: Proinflammatory cytokines in the olfactory mucosa result in covid-19 induced anosmia. ACS chemical neuroscience **11**(13), 1909–1913 (2020)
41. Turjya, R.R., Khan, M.A.A.K., Mir Md. Khademul Islam, A.B.: Perversely expressed long noncoding rnas can alter host response and viral proliferation in sars-cov-2 infection. Future Virology **15**(9), 577–593 (2020)
42. Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J.: Machine learning algorithm validation with a limited sample size. PloS one **14**(11), e0224365 (2019)

43. Vignolo, L.D., Milone, D.H., Scharcanski, J.: Feature selection for face recognition based on multi-objective evolutionary wrappers. Expert Systems with Applications **40**(13), 5077–5084 (2013)
44. Wilson, J.C., Kealy, D., James, S.R., Plowman, T., Newling, K., Jagger, C., Filbey, K., Mann, E.R., Konkel, J.E., Menon, M., et al.: Integrated mirna/cytokine/chemokine profiling reveals severity-associated step changes and principal correlates of fatality in covid-19. Iscience p. 103672 (2021)
45. Xu, J., Zhou, X., Wang, J., Li, Z., Kong, X., Qian, J., Hu, Y., Fang, J.Y.: Rhogaps attenuate cell proliferation by direct interaction with p53 tetramerization domain. Cell reports **3**(5), 1526–1538 (2013)
46. Yan, J., Li, P., Gao, R., Li, Y., Chen, L.: Identifying critical states of complex diseases by single-sample jensen-shannon divergence. Frontiers in oncology **11**, 1824 (2021)
47. Zhang, L., Fang, W.J., Zhang, K.M., Jiang, W.W., Chen, M., Liao, W.Q., Pan, W.H.: Long noncoding rna expression profile from cryptococcal meningitis patients identifies dpy19l1p1 as a new disease marker. CNS Neuroscience & Therapeutics **25**(6), 772–782 (2019)
48. Zhang, T.: Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Twenty-first international conference on Machine learning - ICML '04. ACM Press (2004). https://doi.org/10.1145/1015330.1015332, https://doi.org/10.1145/1015330.1015332
49. Zhou, Z., Li, S., Qin, G., Folkert, M., Jiang, S., Wang, J.: Multi-objective-based radiomic feature selection for lesion malignancy classification. IEEE Journal of Biomedical and Health Informatics **24**(1), 194–204 (Jan 2020). https://doi.org/10.1109/jbhi.2019.2902298, https://doi.org/10.1109/jbhi.2019.2902298