Hybrid Evolutionary Framework for Selection of Genes Predicting Breast Cancer Relapse

Lorenzo Perino
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Milan, Italy
lorenzo.perino@mail.polimi.it

Silvia Cascianelli
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Milan, Italy
silvia.cascianelli@polimi.it

Marco Masseroli
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Milan, Italy
marco.masseroli@polimi.it

***Abstract*—Predicting relapse events is still one of the major challenges for breast cancer research. Despite gene expression-based classifiers may tackle this task, working on thousands of genes and only few samples jeopardizes the performances of a classifier trained without a proper gene selection. We propose a novel hybrid evolutionary gene selection framework, which uses a Multi-Objective Genetic Algorithm (MOGA) to search a wider range of gene selections and handles MOGA results in a whole new way, so as to overcome the limit of the non-easy interpretability of the MOGA broad set of solutions. To a classifier our framework provides a gene signature not only bringing the best cross-validation result, but also having noteworthy and robust performances when tested on unseen samples of an hold-out set. The robustness in hold-out showed the strength of our innovative key element: the final module of the framework, which fully exploits the high variability of MOGA outputs, rather than choosing just one of the solutions, as commonly done in the literature. It combines all MOGA results in more robust and compact gene occurrence-based signatures, under the reasonable assumption that highly recurrent genes have a more crucial biological role, more suitable clinical application and good discriminative power between relapsed and relapse-free patients, as confirmed by the obtained classification results.**

***Index Terms*—Breast cancer relapse; Gene selection; Hybrid feature selection; Multi-Objective Genetic Algorithm**

I. Introduction

Advances in screening and treatment for Breast Cancer (BRCA) have dramatically improved survival rates over the last decades. Nonetheless, BRCA is a complex disease whose risk of recurrence varies greatly from case to case depending on molecular traits, stage at diagnosis and treatments [1]–[3]. Genomic investigation can shed some light to distinguish favourable from risky long-term clinical outcomes, and help physicians in deciding whether or not a patient should be treated with chemotherapy after surgery. Using genomic data to predict relapse events for BRCA patients after primary tumor removal is an important and challenging goal, still to be reached despite being addressed by multiple research groups over the last decades [4]–[6]; they focused primarily on investigating gene expression data, i.e., quantitative measurements of gene activity in specific samples and conditions.

This work was supported by the ERC Advanced Grant 693174 "Data-Driven Genomic Computing (GeCo)".

Microarray technology is widely used to provide gene expression data. Collected datasets are typically characterized by a relatively small number of samples, but a huge number of features, i.e., the thousands of profiled genes. Thus, a proper gene selection phase is crucial to identify significant genes for a predictive task, while discarding irrelevant, redundant or noisy genes. This allows facing both the overfitting risk – i.e., avoiding to learn the noise within the data and to lose generalization capabilities – and the curse of dimensionality – i.e., preventing the predictive power of the learned model trained on a fixed number of samples from decreasing due to the huge and heavily unbalanced feature size.

Feature selection methods typically belong to three general categories: *filters*, *embedded regularizations* and *wrapper techniques*. For intrinsically complex, high and heavily unbalanced dimensionalities, as in gene expression datasets, a fourth category has been emerging: *hybrid methods*; they combine several feature selection and optimization approaches to speed up and improve gene selection. In this trend, genetic algorithms (GA) have become very popular optimization techniques dealing with gene expression data (e.g., see [7], [8]), particularly Multi-Objective Genetic Algorithms (MOGA) [9]–[13], which can simultaneously optimize conflicting objectives and find a set of relevant solutions. This enables comparing more and different solutions from a wider range of evaluations, without severely impact the computational costs.

In this work, we developed a hybrid feature selection framework, combining proven and innovative elements, in order to select gene signatures able to improve the discriminant power between cancer-relapsed and relapse-free patients of binary classifiers working on microarray gene expression data. Notably, we propose a tripartite framework integrating 1) a filter method based on a Signal-to-Noise Ratio (SNR) metric, 2) a wrapper method using a MOGA and a regularized classifier to minimize both gene signature size and classification error, and 3) a gene occurrence-based selection method; this latter one is thought to cope with the high output variability of the previous layer, by progressively grouping genes of final MOGA populations in signatures, from the most occurred ones until the union of all selected genes. It also merges results from multiple MOGA runs to improve robustness of the solutions,

and provides a clear interpretation of the best found signature.

The rest of this paper is organized as follows. Section II is devoted to related works, while Section III to materials and methods. Section IV describes the proposed hybrid framework. In Section V we analyze and discuss the results of our framework and the improvements in predicting BRCA relapses when using the emerged gene signatures. Eventually, conclusions and future developments are reported in Section VI.

II. Related Works

Dealing with gene expression data, different feature selection approaches have been proposed to select task-related gene signatures for further sample classification. Used approaches range from fast filterings, as in [14], [15], to hybrid approaches, usually combining a filter and a wrapper method. Wrapper methods require suitable strategies to ease their search process without heavily penalizing the overall investigation. To this end, greedy heuristics have been used, such as the sequential forward selection, as in [16], or the recursive backward elimination, as in [17]. However, forward selection often converges fast to non-robust results, while backward elimination requires a high computational time for wide gene sets. Both need multiple runs with gene shuffling to face the limit of their sequential scrolling.

Alternatively, some hybrid approaches include genetic algorithms to limit the computational costs while evaluating a broader space of possible solutions, as in [7], [8]. Particularly, MOGA approaches have recently raised an increasing interest in gene expression-based classification or clustering tasks [10]–[13], since gene selection problems can be approached from a multi-objective perspective, typically maximizing a performance measure while minimizing the amount of genes retained. Most recent MOGA related works (e.g., [11]–[13]) are inspired by the so called second generation of MOGA [18], which introduced the concept of elitism. Elitism allows to retain nondominated individuals over generations, so as to assure convergence towards optimality. Among the second generation of MOGA, three algorithms are the most representative: Strength Pareto Evolutionary Algorithm (SPEA) [19], Pareto Archived Evolution Strategy (PAES) [20], and Nondominated Sorting Genetic Algorithm II (NSGA-II) [21]. The two contributions most relevantly related to our work are discussed below.

In 2008 Mohamad *et al*. [10] introduced a MOGA approach based on Support Vector Machines (MOGASVM) to tackle gene selection. They aimed at minimizing the cardinality of the selected gene subset while improving the classification accuracy on validation data. Yet, contrarily to Pareto-based approaches, they weighted the two objective functions, to tune their relative importance, and linearly combined them, so as to obtain a single objective function. Clearly, the downside of this approach lies in returning a single optimal solution rather than a variety of nondominated solutions, as in Pareto-based approaches. Furthermore, assigning weights to objective functions is contradictory to the concept of Pareto-optimality, where same priority is given to all objectives.

In 2016 Hasnat and Molla proposed an hybrid method combining a Correlation Coefficient-based filtering and NSGA-II [12] to select a minimal set of non-redundant genes with the highest classification accuracy from three well-known cancer datasets (about Leukemia, Lymphoma, and Colon cancer). The authors employed a k-nearest neighbors (k-NN) classifier with a leave-one-out cross-validation (LOOCV) scheme. Although obtaining remarkably high classification accuracy, only the highest-accuracy signature is retained from the whole set of MOGA nondominated solutions, without fully exploiting the advantage of a Pareto-based approach such as NSGA-II. Moreover, as pointed out by the authors themselves, using a parametric classifier, like SVM, can be a future enhancement.

In conclusion, there are many examples of using hybrid approaches with MOGA for gene selection in the literature. Nonetheless, to the best of our knowledge, most share two common drawbacks: 1) using classification performance metrics that are not equally sensitive to all types of classification errors, and 2) selecting one best solution out of the nondominated set rather than strive to comprehensively exploit all retrieved solutions to produce more robust results.

III. Materials and methods

A. Dataset used

We used the public dataset, available at NCBI/Genbank GEO database (series entry GSE2034), first employed in Wang *et al* [6]. It includes the expression profiles of 22,283 genes in 286 samples, annotated with relapse events, of patients with lymph-node-negative primary BRCA who did not receive neoadjuvant or adjuvant therapy. These patients were observed for a 5-year post-operative follow-up period to assess their clinical outcome as cancer-relapsed or relapse-free patients.

Gene expression had been measured with Affymetrix oligonucleotide microarray U133a GeneChip. Genes with average expression intensity less than 40 units, or background signal more than 100 units, were excluded. For chip normalisation, probe sets were scaled to a target intensity of 600 units. For our work, gene expression values in the dataset were preliminarily log2-scaled, upper quartile normalized (by sample) and standardized as z-scores. Then, samples were randomly split into training (228 samples) and hold-out (58 samples) sets, with sample allocation stratified by class (relapsed or relapse-free patients) to ensure the same class proportion as in the whole dataset (37% and 63%, respectively).

B. Feature selection methods

We used a cascade of three feature selection methods, to exploit all their advantages and overcome their individual limits, still avoiding unaffordable computational costs.

We adopted an initial filter method, since filtering is a computationally effective pre-processing step that easily scales to high-dimensional datasets. Specifically, genes are ranked according to a statistical scoring function known as Signal-

to-Noise Ratio (SNR). In our relapse (r) and relapse-free (f) two-class context, it gives the following value to each gene g:

$$P_{SNR}(g) = \left|\left| \frac{\mu_r(g) - \mu_f(g)}{\sigma_r(g) + \sigma_f(g)} \right|\right| \quad (1)$$

where μ_r and σ_r are the mean and standard deviation of the expression values of gene g for samples belonging to the class r, while μ_f and σ_f refer to samples of class f. High ranked genes are then kept as feature space of interest, regardless of the model chosen to perform the predictive task and without considering any relationships among features. Hence, selecting the SNR-based top ranked genes helps in preserving the genes with the highest expression variability between the two classes of interest and with the minimal expression variation within each class, while not assuming equality of standard deviations.

Additionally, embedded regularizations are used in each assessed classifier, to learn which genes best contribute to its performance while it is being fitted. Given a parametric model m having θ as parameter vector and $\mathcal{L}_m(\theta)$ as loss function, an embedded regularization introduces an additional term $\mathcal{L}_{reg}$ weighted by its hyperparameter(s) γ, such that the total loss function to be minimized takes the form:

$$\mathcal{L}(\theta) = \mathcal{L}_m(\theta) + \gamma \mathcal{L}_{reg}(\theta) \quad (2)$$

Notably, we added the L_2-regularization term, $\lambda_2 ||\theta||_2$; this shrinks parameter values to decay towards zero, although none is nullified nor the corresponding feature discarded. Differently from L_1-regularization ($\lambda_1 ||\theta||_1$) or Elastic Net (which combines both L_1 and L_2 regularizers), L_2-regularization cannot induce sparsity in a model; thus, we chose this regularizer only to allow even more complex models to be trained over our limited sample size, but high gene dimensionality, without severe overfitting. Setting properly the hyperparameter λ_2 is crucial for the good training of each model; yet, this tuning is worth to improve generalization capabilities.

Lastly, we used a wrapper method, where different gene combinations are generated and compared based on the performances reached by the classifier under evaluation. Specifically, we used a MOGA, rather than a greedy algorithm, to lead a search process where both gene signature sizes and learner performances are optimized over time. In wrapper methods, selected genes are tailored for the considered learner, but computational costs often become prohibitive, since the search space grows exponentially with the number of starting features. Conversely, our MOGA considers a fixed, but reasonably wide and varied, amount of combinations in the search space, overcoming this issue.

C. Classification models

We explored two traditional supervised learning approaches to achieve our binary classification task, while showing the predictive performance improvements provided by the proposed feature selection framework: *Logistic Regression* (LR) and *Support Vector Machines* (SVM). Both have been extensively used for the classification of gene expression data (e.g. in [15], [17]); as any parametric supervised model, they learn their parameters based on training sample-class pairs, as to capture relationship functions from known examples.

Logistic Regression is a classification method that uses the logistic sigmoid function σ on a linear combination of features, weighted by the parameter vector, to estimate the posterior probability of a sample to belong to a class. This simple approach is thought for binary classification, where the alternative class probability is just the complement of the found probability. Specifically, LR minimizes the following loss function:

$$\begin{aligned} \mathcal{L}(\theta) = &- \sum_{i=1}^{n} y^{(i)} \log \sigma\left(\theta^\top x^{(i)}\right) + \\ &+ \left(1 - y^{(i)}\right) \log\left[1 - \sigma\left(\theta^\top x^{(i)}\right)\right] \end{aligned} \quad (3)$$

where σ is the sigmoid function, n is the number of samples, θ is the parameter vector and $y^{(i)}, x^{(i)}$ are the class and feature vectors of the i^{th} sample, respectively.

SVM is originally designed for binary classification tasks on linearly separable data; among the infinite hyperplanes, i.e., the possible linear boundaries able to separate the data, SVM finds the optimal hyperplane that maximizes the margin from the nearest points of each class. In broader terms, SVM can use a kernel function $\mathcal{K}$ to non-linearly transform the data into a higher dimensional space in which data are separable. Nevertheless, not every dataset is separable, even after kernel transformation, and consequently SVM can be reformulated to classify also a complex and noisy dataset by defining a non-perfectly separating hyperplane that anyway minimizes also the classification error.

When true class labels are represented as target values $y^{(i)}$ in the {-1,1} set and $\mathcal{S}$ is the set of the indexes of the support vectors x_s with their associated parameters α_s and bias b, the class prediction $\hat{y}$ for an unseen sample $x^{(u)}$ is computed as:

$$\hat{y}\left(x^{(u)}\right) = \mathbf{sign}\left(\sum_{s \in \mathcal{S}} \alpha_s y^{(s)} \mathcal{K}\left(x^{(u)}, x^{(s)}\right) + b\right) \quad (4)$$

Selecting the best kernel function $\mathcal{K}$ alongside with all the hyperparameters can be challenging during the model selection phase. Notwithstanding, to provide a broader investigation we assessed the performances of all the following kernels: *linear*, *radial basis function* (*RBF*), *sigmoid*, and *polynomial* with degrees from 2 to 5. For each of the corresponding SVM models, we used the *hinge loss*, i.e., $\mathcal{L}(\hat{y}) = max\left(0, 1 - \hat{y}y\right)$, as function to be minimized. Notably, for the SVM with linear kernel we considered also the *squared hinge loss*, where larger errors are punished more significantly than with the traditional hinge function, whereas smaller errors are punished slightly less due to the square of the output of the hinge. Eventually, for all the considered LR and SVM classifiers we used the L_2-regularized models implemented in the scikit-learn Python package [22], as to limiting overfitting risk.

D. Performance metrics

Each classifier was evaluated based on the following performance metrics computed after the model training process, i.e.,

Matthews Correlation Coefficient (*MCC*), *Balanced Accuracy* (ACC_b) and *F1-score*:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5}$$

$$ACC_b = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{6}$$

$$F1{-}score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{7}$$

where TP, TN, FP and FN are true positive, true negative, false positive and false negative values, respectively, while, for the positive class, Precision is $\frac{TP}{TP+FP}$ and Recall is $\frac{TP}{TP+FN}$.

Notably, MCC was used also to compute the classification error within the MOGA of our wrapper method (it is one of the fitness functions to be minimized). MCC is indeed a reliable and symmetrical statistical rate for binary classifications [23]; it gives a high positive score only if the prediction has good results in all four confusion matrix categories (TP, TN, FP and FN), despite the possibly unbalanced sizes of the two classes.

E. Multi-objective optimization using genetic algorithms

Gene selection can be modelled as a multi-objective problem (MOP), given the contrasting objectives involved. On one hand, the cardinality of the subset of selected genes should be minimized in order to ease the assessment of the functional interdependence between genes, and to remove redundant and noisy ones. On the other hand, excessive cardinality reduction leads to a deterioration in classification performance, which is rather to be maximized.

Formally, MOPs with k objective functions can be formulated as finding a candidate solution vector of decision variables $\bar{x}^* = [x_1^*, x_2^*, ..., x_n^*]^T$ satisfying m inequality constraints

$$g_i(\bar{x}) \geq 0, i = 1, 2, ..., m \tag{8}$$

and p equality constraints

$$h_i(\bar{x}) = 0, i = 1, 2, ..., p \tag{9}$$

and optimizes the vector function

$$\bar{f}(\bar{x}) = [f_1(\bar{x}), f_2(\bar{x}), ..., f_k(\bar{x})]^T \tag{10}$$

where the constraints given in (8) and (9) define the feasible region F that contains all admissible solutions.

Hence, a MOP admits multiple solutions representing the best possible trade-offs between the objectives to be optimized; yet, as mentioned, the objectives are often intrinsically conflicting, which poses the issue of how the trade-off among them should be assessed. In the literature, three alternative types of approaches have been reported, i.e., aggregation-based approaches, non-Pareto approaches (such as the lexicographical one), and Pareto-based approaches. The first two types are conceptually straightforward and remarkably easy-to-use; yet, they share the main drawback of returning a single optimal solution, obtained by optimizing a scalar function that linearly combines the objectives. By contrast, although at cost of higher computational complexity, Pareto-based approaches assign equal priority to all objectives involved, keeping the variety of trade-off solutions intact, as to offer the possibility to explore the nondominated solutions and their interdependence. Thus, Pareto-based approaches look more suitable to high dimensional problems such as gene selection, due to their broader exploration of the search space.

However, multiple optimal solutions in the sense of Pareto optimal are admissible in MOPs; they are nondominated solutions, meaning that at least one of their objective function is optimal as to any other feasible solution. Formally, a candidate solution x_1 is a dominating solution to x_2 if:

$$\begin{aligned} &f_i(x_1) \leq f_i(x_2) \,\forall\, i \in 1, 2, ..., k \;\wedge \\ &\exists i \in 1, 2, ..., k \;:\; f_i(x_1) < f_i(x_2) \end{aligned} \tag{11}$$

Hence, a feasible solution is a Pareto optimal solution if it is not dominated by any other feasible solution. When the set of Pareto optimal solutions are mapped in the objective space they are collectively known as Pareto Front. Obtaining the True Pareto Front of a MOP is the ultimate goal of multi-objective optimization algorithms. However, usually only a sub-optimal approximation of the True Pareto Front is achievable due to lacking of a-priori knowledge of the problem.

Likewise, in high dimensional feature spaces, an exhaustive search of all possible solutions is not feasible. Thus, Multi-Objective Genetic Algorithms offer an heuristic approach to perform the search by evolving a population of candidate solutions in a reasonable computational time. For this study, we adopted NSGA-II, a well-known Pareto-based MOGA. NSGA-II makes use of Elitist Selection [12], [21] to prevent the quality of the obtained solutions from degrading over generations. This kind of selection, while constructing a new population, preserves the best individuals from the current generation to the next generation in unaltered form. Furthermore, NSGA-II has been proven to converge to the global Pareto Front while maintaining the diversity of the population, by means of a measure called *crowding distance*. For each Pareto Front solution, the crowding distance operator assigns the highest value to the boundary solutions, thus favoring isolated solutions to be retained over generations.

IV. Proposed hybrid evolutionary framework

The hybrid evolutionary framework we propose for gene selection is characterized by the tripartite structure depicted in Figure 1, whose main components are following described.

A. Preprocessing with filter method

The first component of our framework is a canonical filter method, thought to reduce efficiently the dimensionality of the feature space of interest. We chose a filter based on Signal-to-Noise Ratio, as described in Section III-B, to preserve genes with greater expression variability between the two considered classes and with minimal expression variation within each class, keeping the top 5% genes as filtered gene set of interest.

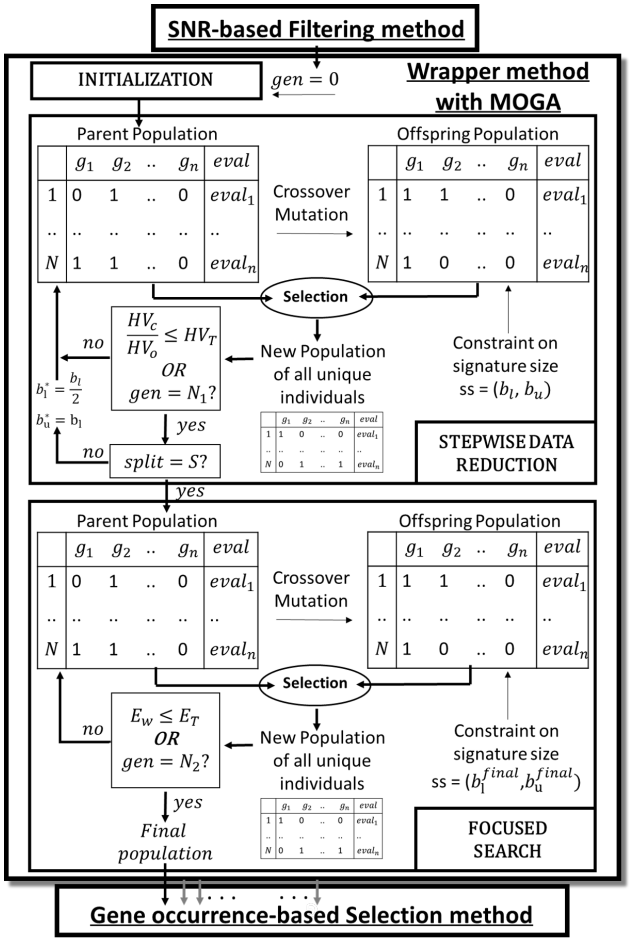

Fig. 1. Hybrid evolutionary gene selection framework

B. Wrapper method with MOGA

The second component of the proposed selection framework is a wrapper method using a MOGA as heuristics. Here, we address the gene selection problem as a MOP, aiming at minimizing simultaneously gene signature size and classification error of the classifier having the signature as the feature space.

We encoded our MOP using fixed-length binary strings $x = \{0,1\}^n$ (called individuals) to represent the gene signatures under evaluation, all of length n. Being n equal to the number of pre-filtered genes, each string binary cell has a one-to-one correspondence with a gene. Thus, the value of a cell equal to zero encodes the absence of the associated gene from the resulting signature, while if it is equal to one the associated gene is included. Finally, the two objective functions to be minimized can be formalized as follows:

$$f_1(x) = count(x_i | x_i = 1)$$

$$f_2(x) = \begin{cases} 1 & \text{if } MCC(x) \leq 0 \\ 1 - MCC(x) & \text{if } MCC(x) > 0 \end{cases}$$

where $f_1(x)$ returns the cardinality of the gene signature x, while $f_2(x)$ is the classification error computed from MCC. Notably, the $MCC(x)$ error of a single individual comes from a 5-fold cross-validation procedure; thus, it is computed as the mean error across five folds.

By means of MOGA, multiple gene signatures are extracted from the whole set of pre-processed genes. The search process starts with the *initialization* of a fixed-size population of individuals, each of them encoding a gene signature. Individuals then evolve over generations and only the fittest individuals are retained by selection.

One of the main innovation of this study lies in the search phase, which progressively shrinks the search space from large to reduced signature sizes (*Stepwise Data Reduction*), as to avoid quick convergence and loss of putative important genes in the initialization. Accordingly, signature size of the offspring individuals is bounded within a range of max b_u and min b_l genes; search is performed over generations within the same range until an update criterion is met, causing the lower bound b_l to be halved and fixing the upper bound b_u equal to the old lower bound value. The criterion for range updating is based on the *hypervolume*, a metric to measure population convergence towards the optimal objective space region [24]. If the hypervolume ratio between the current generation and the first generation after each update, defined as $\frac{HV_c}{HV_0}$, is less than or equal to a preset threshold HV_t, or if a preset number of generations N_1 is reached, the current range of interest is updated. Hence, the search process iterates likewise until the preset number of split updates S is reached. Notwithstanding, over-sized nondominated individuals are preserved over generations (in parent populations) due to elitsm.

Following (*Focused Search*), while keeping a fixed final range of signature sizes $(b_l^{final}, b_u^{final})$, individuals are improved over generations until the classification performance of the worst Pareto Front individual reaches a prefixed threshold E_w, or a preset number of generations N_2 is over.

The output of a single run of the MOGA is its final population, whose complete genetic composition is analysed in the final step of our framework and possibly merged with final populations from other runs, as to increase the robustness of the obtained final gene occurrence-based signatures.

C. Gene occurrence-based selection

This last component of our framework provides a complete analysis of the genes in final MOGA populations. Given a set S of signatures s_i constituting a final MOGA population, first the whole genetic pool u_{genes} of S is retrieved as those genes occurring in at least one signature s_i. Subsequently, genes are ranked based on the occurrence occ_j of each gene j in the set of final signatures S; then, they are progressively included in gene occurrence-based signatures (gobs). The inclusion procedure is such that for each threshold occ_t, the correspondent gobs includes all genes whose occurrence occ_j is greater than or equal to occ_t. The threshold value is lowered unit-by-unit so as to generate gobs for all inclusion levels, from the ones including only most recurrent genes to the one including all genes in u_{genes}. Eventually, from the whole set of gene occurrence-based signatures, the one bringing the best classification performances in cross-validation is selected as feature space to improve the learner under exam.

V. Results and discussion

A. Gene filtering

The Signal-to-Noise Ratio-based filtering, applied on the initial whole dataset of 22,283 genes, returned 1,115 genes whose expression values are the most differentiated between the two classes of cancer-relapsed or relapse-free patients.

B. Model selection

During model selection phase, we compared several classification models with respect to our binary classification task: a Logistic Regression, a SVM with linear kernel and squared hinge loss, and other SVMs with hinge loss and alternatively linear kernel, RBF kernel, sigmoid kernel, or polynomial kernel. We used the SNR-based filtered set of 1,115 genes as starting feature space to investigate their performances, and selected the most promising models as candidates to be further improved in the additional layers of our framework.

Each model was trained using embedded L_2-regularization to prevent overfitting, without inducing further sparsity beyond the gene selection provided by our framework. To set all the hyperparameters of each classifier according to our predictive task, model selection was performed using hyperparameter grid search in combination with a stratified 5-fold cross-validation (CV); the average CV perfomances over the five folds were computed in terms of MCC for all models and were responsible for the final choice of the hyperparameters. Following, the best hyperparameter setting of a model was used to re-train the same model on the whole training set of 228 samples, before assessing its classification results over the 58 unseen samples of our hold-out (HO) set.

The performances of all models under evaluation are reported in Table I, where LR and SVM with RBF kernel emerge as the most favourable classifiers. In a supplementary analysis, we assessed also the average performances (in 10-fold CV, 5-fold CV and HO) of all considered models while varying the training and hold-out compositions, but keeping the same proportion between them and the same stratification of relapsed and relapse-free patients, as in the entire cohort. This additional analysis (data not shown) confirmed the goodness and robustness of the performances for the previously emerged LR and SVM with RBF kernel, besides adding another classifier valuable of further investigations: the linear SVM with squared hinge loss. These three models (hereafter called baselines) were hence selected as alternative learner within our whole framework, with the aim of improving their classification performances.

TABLE I
Performances of the classifiers trained using the 1,115 genes from the SNR-based filtering

Assessed Model	Loss Function	5-f CV	HO scores		
		MCC	***MCC***	***ACC$_b$***	***F1***
LR	Cross-entropy	0.533	0.684	0.852	0.808
SVM ($\mathcal{K}_{linear}$)	Squared Hinge	0.522	0.596	0.807	0.760
SVM ($\mathcal{K}_{linear}$)	Hinge	0.522	0.596	0.807	0.760
SVM ($\mathcal{K}_{RBF}$)	Hinge	0.542	0.704	0.845	0.810
SVM ($\mathcal{K}_{poly}$)	Hinge	0.311	0.474	0.659	0.483
SVM ($\mathcal{K}_{sigmoid}$)	Hinge	0.524	0.628	0.799	0.750

C. MOGA and gene occurrence-based selection

The SNR-based filtered dataset fed to our proposed MOGA contained 228 samples and 1,115 genes. Twenty different runs for each of the 3 classifiers emerged from model selection were performed, taking each time a population of 100 individuals with length = 1,115, uniform crossover probability = 0.9, and bit-wise mutation probability = $\frac{1}{1,115}$. In the Stepwise Data Reduction phase, the range of admissible values for the signature size was updated four times (S = 4). The lower b_l and upper b_u boundaries were sequentially updated as follows: $[1, 115, 557]$, $[557, 278]$, $[278, 139]$ and $[139, 69]$. The threshold for the hypervolume ratio HV_t was set to 0.5, while the maximum number of generations N_1 within each subset size range was set to 100. As for the Focused Search phase, the fixed range of admissible subset sizes was set to $[69, 34]$. For the two stopping criteria, the worst error threshold of the Pareto Front E_t was set to 0.4, while the maximum number of generations N_2 to 200. All these values were heuristically chosen while balancing computational time and search width.

As a common trend among all MOGA runs, the MOGA population was able to half the initial hypervolume HV_0 within the maximum number of generations allowed for each subset size range explored, confirming the high convergence rate characteristic of NSGA-II. Equal behavior appeared in the Focused Search phase, where, regardless of the classifier, all runs satisfied the stopping criterion on the worst error threshold E_t before reaching the maximum number of generations

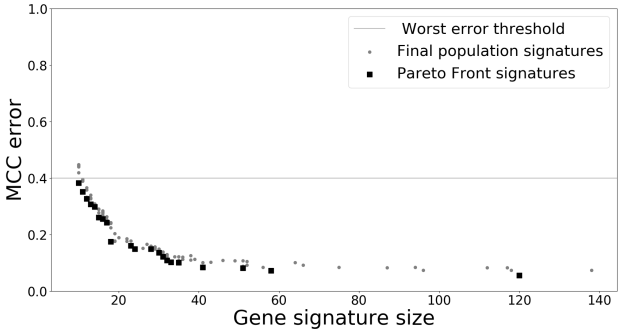

Fig. 2. A final MOGA population with its Pareto Front in the objective space

Figure 2 exemplifies the final population obtained after a single run of MOGA, with its Pareto Front highlighted. Notably, signatures of the final population may not fall in the last subset size range $[69, 34]$, since final individuals are the result of the Elitist Selection across all explored ranges. Nonetheless, the Pareto Front shows higher density of non-dominated solutions in the aforementioned last range, being the final search process focused in that region.

For each classifier of interest, the last layer of our framework merges the final populations from twenty MOGA runs, each with a different random initialization, into a unique population, denoted hereafter as U_{20} (population size = 2,000

gene signatures). Then, it retrieves the genetic pool across all signatures in U_{20}, i.e., the union of the genes occurring in at least one signature in U_{20}, hereafter named as u_{genes}. Eventually, the genes in u_{genes} are sorted in descending order based on their occurrences in U_{20}, and gene occurrence-based signatures are composed as described in Section IV-C.

Assessment of fitness for the gene occurrence-based signatures was performed as for the MOGA populations, by computing their size and $MCC\ error$ in 5-fold cross validation. Moreover, these signatures were adopted as feature space to re-train the corresponding classifier on the whole training set, with the best hyperparametrization emerged from cross-validation. Each classifier was then used to predict the unseen samples of the hold-out set, to test its generalization property. As an example, results of linear SVM are shown in Figure 3.

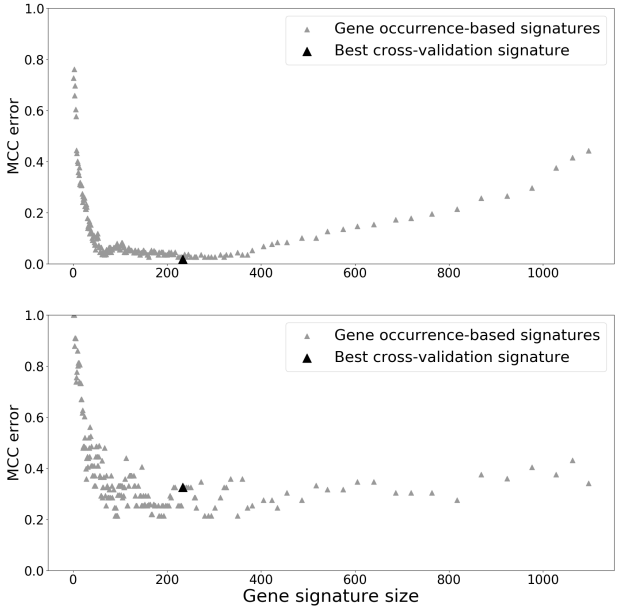

Fig. 3. Gene occurrence-based signatures of linear SVM, in cross-validation (above) and hold-out (below) objective spaces

Gene occurrence-based signatures with a reduced size (i.e., including only highly occurred genes) do not achieve good classification performances neither in CV or HO; same trend is observed for large-sized signatures including also rarely occurred genes. However, signatures with low classification error in CV are able to replicate overall good performances when predicting HO unseen samples, suggesting that they include discriminative genes across the CV and HO sets.

Lastly, for each of the three classifiers, the gene occurrence-based signatures with the best CV performances were extracted and their HO prediction results compared to the absolute best results reached in HO, as reported in Table II. Overall, best CV signatures perform closely in HO to the best results experienced in HO, as their difference (ΔMCC) stays restrained. Furthermore, their performances in HO are much better than the ones reached with the best CV signatures found with MOGA (data not shown); indeed, the innovative last module of our framework brings a remarkable improvement of MCC (up to 0.150).

TABLE II
GENE OCCURRENCE-BASED SIGNATURES RESULTING FROM THE UNION OF 20 MOGA RUNS FOR THE 3 CLASSIFIERS SELECTED

Assessed Classifier	Signature			5-f CV	HO Scores	
	size	size%	occ_t%	*MCC*	*MCC*	Δ*MCC*
LR	151	13.93	≥ 2.65	0.973	0.611	0.208
	196	18.08	≥ 1.9	0.973	0.746	0.073
	228	21.03	≥ 1.6	0.973	0.674	0.145
	246	22.69	≥ 1.4	0.973	0.707	0.112
SVM ($\mathcal{K}_{linear}$)	233	21.24	≥ 2.05	0.982	0.674	0.112
SVM ($\mathcal{K}_{RBF}$)	69	6.28	≥ 9.4	0.963	0.741	0.076
	71	6.47	≥ 8.55	0.963	0.705	0.113

Intriguingly, for Logistic Regression, signatures different in size (and size[%], as percentage of u_{genes} included) and gene occurrence threshold (expressed in $occ_t[\%]$ as percentage of total MOGA signatures found) share the same optimal classification performance in 5-fold CV. Conversely, their HO performances are largely different from one another (HO Scores - MCC column in Table II), suggesting the need for LR of further MOGA runs to achieve more robust results.

D. Comparative evaluation

Classification results of each baseline were compared with the ones of the same classifier when trained with only the signature emerged from our gene selection framework. When multiple gene occurrence-based signatures resulted optimal in CV, we selected the largest signature in size as comparison term for each classifier. This is a conservative choice due to the supposedly higher generalization capability of a more inclusive signature, yet not observed in this study. Performance in predicting the HO set were compared not only on the basis of the resulting MCC, but also adopting three more metrics: balanced accuracy (ACC_b) and F1-score assessing either relapsed or relapse-free predictions (named $F1_+$ and $F1_-$, respectively), as illustrated in Table III.

Our proposed gene selection framework was able to improve the baseline score of all considered classifiers, with a noteworthy enhancement for Logistic Regression and SVM ($\mathcal{K}_{linear}$), the former one resulting the best overall. Although performances for the SVM ($\mathcal{K}_{RBF}$) classifier remained almost unchanged, the gene occurrence-based signature eventually selected shows a robust performance in HO (MCC = 0.705), while being remarkably reduced in size (71). This latter achievement in particular makes the signature more suitable

TABLE III
PERFORMANCE EVALUATION BEFORE AND AFTER OUR GENE SELECTION

Assessed Classifier	Genes of Interest	HO Scores			
		MCC	ACC_b	$F1_+$	$F1_-$
LR	1,115	0.684	0.852	0.808	0.844
	246	0.707	0.854	0.818	0.889
SVM ($\mathcal{K}_{RBF}$)	1,115	0.704	0.845	0.810	0.892
	71	0.705	0.845	0.811	0.890
SVM ($\mathcal{K}_{linear}$)	1,115	0.596	0.807	0.760	0.819
	233	0.674	0.840	0.800	0.873

to clinical application, as opposed to the baseline gene set (1,115). Lastly, as expected, both baseline and best gene occurrence-based signatures showed better performances when predicting non-relapsed patients ($F1_-$) as compared to relapsed ones ($F1_+$), being the former one the most represented class in the training set.

VI. Conclusions

To effectively distinguish BRCA relapsed from relapse-free patients, in this work we proposed a hybrid gene selection framework able to find a suitable gene signature for each classifier to accomplish the task. The framework joins proven approaches and innovative elements in a tripartite structure: a filter method based on Signal-to-Noise Ratio (SNR) metric, a wrapper method using a NSGA-II-based customized MOGA together with a regularized classifier (LR or SVM), and a novel gene occurrence-based selection method. Notably, the implemented MOGA minimizes both the gene signature size and the classification error, based, as never before, on the Matthews Correlation Coefficient, an highly reliable metric in binary classification contexts, particularly when classes are unbalanced. A further innovation lies in the adopted search technique, since the search space is spanned by iteratively halving the size of the signatures under exam, as to prevent quick convergence and random initialization from causing the loss of putative task-related genes. However, the last module of our framework is the major strength, since, rather than choosing a single optimal MOGA solution, it fully exploits the high variability of MOGA outputs with meaningful improvements over mere MOGA results. Particularly, it combines all genes emerged from MOGA solutions in more robust, compact gene occurrence-based signatures, easier to interpret and having higher discriminative capabilities. When tested on the unseen samples of the hold-out set, our best classifiers retain robust and valuable performances, in line or even considerably improved compared to the same learner trained over the 1,115 SNR-based filtered genes, with the additional gain of dealing with more compact and easy to interpret signatures. Finally, the biological interpretation from enrichment analysis of the best gobs sees a significant over-representation of biological processes impacting on *immune response* and *cell cycle*.

Acknowledgment

The authors thank T. Hiroyasu and S. Hiwa from Doshisha University (Kyoto, Japan) for supervision on MOGA.

References

[1] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[2] A. Prat and C. M. Perou, "Deconstructing the molecular portraits of breast cancer," *Molecular Oncology*, vol. 5, no. 1, pp. 5–23, 2011.

[3] A. Ignatov, H. Eggemann, E. Burger, and T. Ignatov, "Patterns of breast cancer relapse in accordance to biological subtype," *Journal of Cancer Research and Clinical Oncology*, vol. 144, no. 7, pp. 1347–1355, 2018.

[4] A. F. Vieira and F. Schmitt, "An update on breast cancer multigene prognostic tests — emergent clinical biomarkers," *Frontiers in Medicine*, vol. 5, p. 248, 2018.

[5] P. H. Abreu, M. S. Santos, M. H. Abreu, B. Andrade, and D. C. Silva, "Predicting breast cancer recurrence using machine learning techniques: a systematic review," *ACM Computing Surveys (CSUR)*, vol. 49, no. 3, pp. 1–40, 2016.

[6] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu *et al.*, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[7] P. Singh, A. Shukla, and M. Vardhan, "Hybrid approach for gene selection and classification using filter and genetic algorithm," in *2017 International Conference on Inventive Computing and Informatics (ICICI)*. IEEE, 2017, pp. 832–837.

[8] K. Yan and H. Lu, "An extended genetic algorithm based gene selection framework for cancer diagnosis," in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*. IEEE, 2018, pp. 43–47.

[9] M. Banerjee, S. Mitra, and H. Banka, "Evolutionary rough feature selection in gene expression data," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 4, pp. 622–632, 2007.

[10] M. S. Mohamad, S. Omatu, S. Deris, and M. Yoshioka, "Multi-objective optimization using genetic algorithm for gene selection from microarray data," in *2008 International Conference on Computer and Communication Engineering*. IEEE, 2008, pp. 1331–1334.

[11] G. Chakraborty and B. Chakraborty, "Multi-objective optimization using Pareto GA for gene-selection from microarray data for disease classification," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2013, pp. 2629–2634.

[12] A. Hasnat and A. U. Molla, "Feature selection in cancer microarray data using multi-objective genetic algorithm combined with correlation coefficient," in *2016 International Conference on Emerging Technological Trends (ICETT)*. IEEE, 2016, pp. 1–6.

[13] S. Basu, S. Das, S. Ghatak, and A. K. Das, "Strength Pareto evolutionary algorithm based gene subset selection," in *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. IEEE, 2017, pp. 79–85.

[14] V. Bolón-Canedo, K. Sechidis, N. Sánchez-Marono, A. Alonso-Betanzos, and G. Brown, "Exploring the consequences of distributed feature selection in DNA microarray data," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 1665–1672.

[15] D. A. V. Ca and V. Mc, "Gene expression data classification using support vector machine and mutual information-based gene selection," *Procedia Computer Science*, vol. 47, pp. 13–21, 2015.

[16] M. Abinash and V. Vasudevan, "A hybrid forward selection based lasso technique for liver cancer classification," in *Nanoelectronics, Circuits and Communication Systems*. Springer, 2019, pp. 185–193.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using Support Vector Machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[18] S. Sabzevari and S. Abdullah, "Gene selection in microarray data from multi-objective perspective," in *2011 3rd Conference on Data Mining and Optimization (DMO)*. IEEE, 2011, pp. 199–207.

[19] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 4, pp. 257–271, 1999.

[20] J. D. Knowles and D. W. Corne, "Approximating the nondominated front using the Pareto archived evolution strategy," *Evolutionary Computation*, vol. 8, no. 2, pp. 149–172, 2000.

[21] J. C. H. Hernandez, B. Duval, and J.-K. Hao, "A genetic embedded approach for gene selection and classification of microarray data," in *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Springer, 2007, pp. 90–101.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] D. Chicco and G. Jurman, "The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.

[24] C. M. Fonseca, P. J. Fleming *et al.*, "Genetic algorithms for multiobjective optimization: Formulation discussion and generalization." in *5th Int. Conf. Genetic Algorithms (ICGA)*, vol. 93. ACM, 1993, pp. 416–423.