

# DOSA-MO: Dual-stage Optimizer for Systematic overestimation Adjustment in Multi-Objective problems improves biomarker discovery

Luca Cattelani<sup>\*1</sup> and Vittorio Fortino<sup>†1</sup>

<sup>1</sup>Institute of Biomedicine, School of Medicine, University of Eastern Finland, Kuopio, 70211, Finland.

## Abstract

The challenge in biomarker discovery and validation using machine learning from omics data lies in the abundance of molecular features but scarcity of samples. Most machine learning-based feature selection methods necessitate of hyperparameter tuning, typically performed by evaluating numerous alternatives on a validation set. Every evaluation has a performance estimation error and when the selection takes place between many models the best ones are almost certainly overestimated. Biomarker identification is a typical multi-objective problem with trade-offs between the predictive ability and the parsimony in the number of molecular features. Genetic algorithms are a popular tool for multi-objective optimization but they evolve numerous solutions and are prone to overestimation. Methods have been proposed to reduce the overestimation after a model has already been selected in single-objective problems, but to the best of our knowledge no algorithm existed that was capable of reducing the overestimation during the optimization, leading to a better model selection, or that had been applied in the more general domain of multi-objective problems. We propose DOSA-MO, a novel multi-objective optimization wrapper algorithm that learns how the original estimation, its variance, and the feature set size of the solutions predict the overestimation, and adjusts the expectation of the performance during the optimization, improving the composition of the solution set. We verify that DOSA-MO improves the performance of a state-of-the-art genetic algorithm on left-out or external sample sets, when predicting cancer subtypes and/or patient overall survival, using three transcriptomics datasets for kidney and breast cancer. Since to the best of our knowledge there was no measure of estimation error for multi-objective solution sets, we propose two such measures. According to both of them, DOSA-MO provides more realistic performance estimates in all the considered scenarios.

**Data and source code availability.** The gene expression data used in this study is from public repositories. The source code and detailed numerical results will be available in a public server after peer review.

---

<sup>\*</sup>Corresponding Author: luca.cattelani@uef.fi

<sup>†</sup>Corresponding Author: vittorio.fortino@uef.fi

# 1 Introduction

Molecular biomarker discovery with machine learning (ML) is usually limited by data that includes many features but few samples. This renders the models prone to overfitting and the evaluation prone to estimation error. Many ML approaches involve hyperparameter tuning and feature selection, with model selection based on the performance measured on a validation set within a training, validation, and test paradigm. Comparing many models and choosing the best is prone to overestimation, i.e. the difference between the performance measured on the validation set and the real performance, a.k.a. “winners curse”. The models that fit the noise present in the validation set are advantaged, a phenomena sometimes referred as overfitting on the validation set [1]. This is also exacerbated by data scarcity. In the quest to reach higher accuracies, increasing the number of hyperparameter configurations may lead to selected models with better performance on the validation set, but a corresponding increase in overestimation may determine diminishing or even negative improvements on the test set.

In biomarker selection both the predictive performance (in terms of disease type, survivability, etc.) and the practical applicability are crucial factors. A biomarker that includes a small number of molecular features requires less resources when applied clinically, thus might be preferred over a slightly more accurate but more expensive or time consuming alternative.

Characterizing all the best compromises (or trade-offs) between predictive value and feature set size is a multi-objective (MO) optimization problem [2], and it can be solved by means of MO feature selection (MOFS) techniques [3, 4, 5, 6]. These techniques aim to identify not just a single best solution, as in single-objective (SO) problems, but rather a Pareto front of solutions. This front is the set of optimal solutions that illustrate the trade-offs between different objectives in MO optimization. However, all candidate solutions (or biomarker models selected by the employed MO feature selection technique) are evaluated on the validation set, which can result in the overestimation of the performance of the selected models.

K-fold cross-validation (CV) is the most common methodology for ML assessment and its benefits and limitations have received much attention in the SO case [7, 8, 9]. K-fold CV returns a model trained on all the available samples and an estimation of its performance computed by averaging  $k$  CV results. The obtained model performance is usually underestimated when only one hyperparameter configuration is used. However, it tends to be overestimated when multiple configurations are evaluated, and the model with the best performance is returned [1].

While some methods have been proposed to improve the estimation of the performance when k-fold CV is applied for model selection, their design and application has been limited to SO problems that are a special subset of the more general MO problem. To the best of our knowledge, no attempt has been made to improve the performance estimation in MO problems like the identi-

fication of the best biomarkers in high-dimensional omics data.

MO problems may require more model evaluations when many good trade-offs exist to be discovered. It may be argued that this offers the possibility to model differences in overestimation in function of the position in the non-dominated front.

Limitedly to SO problems, approaches exists to improve the performance estimation of a selected model, but the model is still selected using the unadjusted estimation, so there is no improvement in the model selection. In the model selection process, where solutions are ranked based on a model evaluation metric, simply subtracting a constant from this metric will not affect the ranking and, therefore, the model selection. If instead the adjustment takes into account some characteristic of the models, such as the variance in performance during the validation phase or the feature set size, it might change their order. It may be argued that such an adjustment could improve the model selection if the model selection would use the adjusted estimations.

While the quality of the performance estimation can be easily measured in SO problems as the absolute difference between the performance expected by the algorithm and the performance measured on new samples, to the best of our knowledge no such measure exists for MO problems, where the results are solution sets. Here we propose two new measures for the quality of the performance estimation.

In Cattelani et al., Fig. 1 and Supplementary Fig. 1 are examples of solution set performance, resulting after executing MO optimization for breast cancer subtypes biomarker discovery [5]. The optimizer explored the best trade-offs between balanced accuracy and number of features, and used 3-fold CV internally to evaluate the solutions. The overestimation (difference between the “inner cv” and “test” dots in the figures) of the balanced accuracy increases as the number of features increases. Analogously, the overestimation increases when the estimation increases. Correlations like these may suggest that characteristics of the solutions might be predictive of the overestimation.

To the best of our knowledge, no previous work experimented the effectiveness of methods for mitigating the overestimation in ML applications with 2 or more objectives. Additionally, no previous work applied the adjustment to the performance estimation prior to the selection of the solution set, thus affecting the selection. Consequently, this approach is also yet to be explored in biomarker discovery within large-scale omics data, where adjusting the expected performance of evaluated biomarker models across multi-cohort datasets could lead to a more robust and reliable biomarker selection.

We present a new MO optimization algorithm that wraps any other MO optimizer to produce temporary solutions. These solutions are then employed to train regression models, which are tasked with predicting performance overestimation. Subsequently, the algorithm executes a wrapped MO optimizer once more, but with a key distinction: it now utilizes the regression models to deliver adjusted fitness evaluations, enhancing the model

selection process. The final resulting models are then trained on all the available samples, thus no samples are lost in order to compute their expected performance on new data.

In our benchmarking study, which concentrates on selecting cancer biomarkers from transcriptomic data, we have experimentally verified that incorporating information from the original performance estimation, its variance, and the size of the feature set can enhance both the performance and its estimation. This improvement is evident when predicting cancer subtypes and patient overall survival, particularly on test sets that are either excluded from the original dataset or entirely external.

Tsamardinos et al. [10] compare double CV, the Tibshirani and Tibshirani method [11], and nested CV in their ability to improve the estimation of the fitness for SO problems. These algorithms adjust the fitness estimation but do not affect the model selection, and the model that results is the same that is selected with simple k-fold CV with hyper-parameter optimization.

Automated ML (AutoML) tools try numerous combinations of models and hyperparameters and return the best performing model and an estimation of its performance. In Tsamardinos et al. six AutoML tools are compared [12]. Of these, only one has a predictive performance estimation strategy that adjusts for multiple model validations (limitedly to SO problems), while the majority of the tools have the necessity to withhold a test set for an unbiased estimation of the performance of the winning model, thus loosing samples from the final model training. Differently from our proposed algorithm, none of the cosidered AutoML tools applies a performance adjustment in MO problems, and no considered tool uses the performance adjustment during the optimization for improving the selection of the winning model. To the best of our knowledge our approach is the first that can be applied to the more general MO problems and that improves the selection of the solutions in addition to their fitness estimation.

## 2 Methods

The DOSA-MO algorithm wraps a MO optimizer, serving two main purposes: improving the estimate of the performance of the solutions, and increase the overall performance of the solution set (set of biomarker models in our case study). The data-driven approach complicates the measurement of these achievements, as CV yields varied performance metrics. These include the performance anticipated by the optimizer, based on training data, and the actual performance measured on left-out samples post-optimization. To our knowledge, there is not an established metric for the error in evaluating the predictive performance of solutions in MO problems within a CV framework. In single-objective scenarios, one might simply assess the absolute difference between the fitness expected from training data and the fitness observed on new data. However, in MO scenarios, it is crucial to consider each solution's contribution to the Pareto front. In this section, we introduce two novel met-

rics to measure this estimation error in MO CV setups.

Most importantly, we introduce a novel wrapper algorithm to improve the precision of the fitness estimation and the overall performance in MO optimization and prove its effectiveness in feature selection for biomarker discovery. Our wrapper algorithm follows a two-stage process using inner MO optimizers. In our case study, these optimizers are represented by genetic algorithms (GAs) specifically designed for MO problems and paired with supervised ML algorithms, such as the Gaussian Naïve Bayes (NB) classifier to distinguish different cancer subtypes or the Cox Proportional-Hazards Model (Cox) for survival analysis. In the first stage, regression-based models are trained to estimate the overestimation, drawing on characteristics of the solutions selected by an MO optimizer. Key characteristics include the original performance, its variance, assessed using only training data, and the size of the feature set. Subsequently, in the second stage, these trained regression models are employed to enhance the evaluation of solutions by an inner MO optimizer. This step aims to refine the selection of the solutions, thereby improving the overall results of the process.

We have benchmarked our algorithm, DOSA-MO, in the context of MOFS for cancer biomarker discovery. This was done using gene-based transcriptomic datasets from The Cancer Genome Atlas (TCGA) project [13]. Specifically, our goal is to identify biomarkers for categorizing cancer subtypes and survival prediction in kidney and breast cancer patients through a MOFS approach. For breast cancer, an additional cohort from The Sweden Cancerome Analysis Network - Breast (SCAN-B) [14] serves as an external validation set. Moreover, as inner MO optimizer, we use a novel modification of Non-dominated Sorting Genetic Algorithm III (NSGA3) [15], NSGA3 with clone-handling method and specialized mutation operator (NSGA3-CHS), with NB or Support Vector Machine (SVM) as inner classifiers, and Cox as inner survival model. We compare the unadjusted optimizer with adjustments by 5 different regression models, using both internal-external CV and external validation.

### 2.1 Measuring overestimation in multi-objective problems

We propose two methods to measure the error of the performance estimates for the solutions to MO problems: MO performance error and Pareto delta.

#### 2.1.1 Multi-objective performance error

We define the MO performance error  $E_v$ , starting from the HyperVolume (HV) computed on the train performance ( $H_t$ ) and the Cross HyperVolume (CHV) ( $H_v$ ).  $H_t$  and  $H_v$  have been formally defined by Cattelani et al. ([6]). Since  $H_v$  is a family of functions, with the specific instantiation that depends on the user provided function  $v$ ,  $E_v$  is a family of functions too.

$$E_v(X, X') = |H_t(X, X') - H_v(X, X')| \quad (1)$$

Where  $X$  encodes the performance of the solutions evaluated on the train data, and  $X'$  the performance evaluated on the test data. In our case study we use the same instantiation of  $v$  as in Cattelani et al.: the function  $\lambda$  ([6]).

$E_v$  has a simple definition and can be computed very efficiently if the experimental setup already includes the measuring of  $H_t$  and  $H_v$ . It can be seen intuitively as the difference between the aggregated performance of all the solution set as expected by the optimizer taking into account only train data, and the aggregated performance of the same solutions when applied on never before seen test data by decision makers that must choose their preferred solution informed by train performance only.

### 2.1.2 Pareto delta

The MO performance error is a metric that derives straightforwardly from the train HV and the CHV. This metric does not pinpoint the specific origins of differences between these two measures. Even though the CHV inherently considers the differences between train and test performance inside its formulation, increases in performance in part of the solutions can even out with decreases in another part, leading to a reduced performance error measurement despite these differences between train and test performance that have opposite sign. To address this issue, we introduce an additional measure called Pareto delta. It has the property of being equal to 0 only when there is no difference in performance between train and test data for all elements of the solution set.

To calculate the Pareto delta, we first sum up the absolute error in fitness estimation for each solution, multiplying it by the derivative of the HV concerning that specific objective and solution. Then, we compute the average across all objectives.

Let  $n$  be the number of solutions and  $m$  the number of objectives. Let  $X$  be an  $n \times m$  matrix where  $x_{i,j}$  is the train performance for the  $j^{th}$  objective of the  $i^{th}$  solution, with  $1 \leq i \leq n$  and  $1 \leq j \leq m$ . Similarly, let  $X'$  be an  $n \times m$  matrix where  $x'_{i,j}$  is the test performance for the  $j^{th}$  objective of the  $i^{th}$  solution. Let  $H_t$  be the HV computed on the train performance  $X$ .  $\partial H_t / \partial x_{i,j}$  is the partial derivative of the HV with respect to  $x_{i,j}$ . We define the Pareto delta function ( $P_\Delta$ ) as

$$P_\Delta(X, X') = \frac{1}{m} \sum_{j=1}^m \sum_{i=1}^n |x_{i,j} - x'_{i,j}| \frac{\partial H_t}{\partial x_{i,j}} \quad (2)$$

In the special cases with 0 dimensions or 0 solutions we define  $P_\Delta$  as 0. As long as the differences  $|x_{i,j} - x'_{i,j}|$  are small,  $P_\Delta$  is proportional to the difference between the volume of the geometrical union of the space under the train and test fronts and the volume of the intersection of the same two spaces.

## 2.2 DOSA-MO: Dual-stage Optimizer for Systematic overestimation Adjustment in Multi-Objective problems

The DOSA-MO is a wrapper algorithm that implements a two-phased approach. Initially, regression models are trained to predict overestimation based on selected solutions' characteristics: their the original fitness, its variance and the feature set size in our case study. The original fitness and its variance are general characteristics for solutions to any ML problem, while the feature set size is specific for feature selection solutions. Then, in the subsequent phase, these regression models are used to refine the MO optimizer's solution assessment, focusing on better model selection to enhance overall performance. In more details, DOSA-MO consists of three main steps.

**1. Generating solution sets for overestimation prediction.** It collects solutions to be used as training samples to learn how to adjust the objective functions that are used to evaluate solutions for supervised learning tasks, such as classification accuracy. This consists in running MO optimizers in a k-fold CV loop. For each fold, a solution set is produced using only training data, and its performance is measured on the left-out samples.

**2. Training of regression models for overestimation.** For each objective DOSA-MO trains a regression model on the samples collected during step 1. Each sample contains as independent variables three meta-features of the solutions that are potentially predictive of the overestimation. They are the original fitness (i.e. the fitness used by the optimizer to choose the best solutions, measured using only training data, applying inner k-fold CV in our case study), the standard deviation of that fitness (we measure it using bootstrap), and the number of features included in the solution (number of genes forming the biomarker in our case study). The dependent variable is the overestimation: the difference between the original fitness and the fitness that is computed on new data through CV. Not all solutions can be considered as equally important. We might expect solutions that are in crowded areas of the Pareto front to be selected less often by a decision maker. To take this into account each sample is weighted according to the partial derivative of the HV (constructed using the original fitnesses) with respect to the considered solution and objective. The weights for each fold and objective are scaled to sum to 1. A regression model is trained on these samples for each objective. We minimize the absolute error, when allowed by the specific model, since the impact of an error on the HV is approximately linear for small errors.

**3. Generating the solution set using adjusted performance.** A second MO optimizer is deployed to generate a solution set (each solution refers to a feature set in our case study), with objectives that are systematically adjusted by previously trained regression models for overestimation. In the revised process, each original objective function is replaced by a pipeline that initially computes the function's result, its standard deviation, and the feature count of the solution. This data is then feed to the corresponding adjuster regres-

---

**Pseudocode 1** `MultiObjectiveOptimizer` abstract class definition.

---

```
class MultiObjectiveOptimizer
    method optimize(objectives, trainingData)
```

---

**Pseudocode 2** `DosaMO` class definition.

---

```
class DosaMO
    inherits MultiObjectiveOptimizer

    method new(
        tuningOptimizer, adjusterLearner,
        mainOptimizer):
        self.tuningOptimizer = tuningOptimizer
        self.adjusterLearner = adjusterLearner
        self.mainOptimizer = mainOptimizer

    method optimize(objectives, trainingData):
        foldsData =
            createFolds(trainingData)
        foldHofs = [
            self.tuningOptimizer.optimize(
                objectives, f.train)
            for f in foldsData]
        for i in 1::objectives.size:
            obj = objectives[i]
            weights = [
                assignWeights(
                    h.fitnessHyperboxes(), i)
                for h in foldHofs]
            adjuster = trainAdjuster(
                self.adjusterLearner, obj,
                foldHofs, foldsData, weights)
            adjustedObjectives[i] =
                createAdjustedObjective(
                    obj, adjuster)
        return self.mainOptimizer.optimize(
            adjustedObjectives, trainingData)
```

---

sion model, which predicts the overestimation. The final adjusted performance is calculated by deducting the predicted overestimation from the original fitness value. The solution set generated by the MO optimizer during this final run represents the output of the whole DOSA-MO.

### 2.2.1 Pseudocode formulation of the DOSA-MO algorithm

In order to wrap any MO optimizer, the DOSA-MO must be polymorphic, so we define an abstract class `MultiObjectiveOptimizer` representing a generic MO optimizer (Pseudocode 1). It has a single method `optimize` that works with provided objectives and training data. The DOSA-MO is a `MultiObjectiveOptimizer` itself (Pseudocode 2).

The method `new` is just a simple constructor that saves the polymorphic parts of the algorithm: the `tuningOptimizer`, a MO optimizer used to create the samples for the adjusting regression, the actual regression model (`adjusterLearner`), and the `mainOptimizer` that is the MO optimizer that uses the objectives adjusted by the regressions to produce the results for the user.

The `DosaMO` implementation of `optimize` first organizes the data into folds. The resulting object, `foldsData`, contains the data itself and also the description of how it is

---

**Pseudocode 3** `trainAdjuster` function definition.

---

```
trainAdjuster(
    adjusterLearner, obj, foldHofs,
    foldsData, weights):
    for i in 1::foldsData.size:
        allFitnesses =
            evaluateWithCV(
                foldHofs[i], obj, foldsData[i])
        originalFitnesses[i] = allFitnesses.innerCV()
        stdDevs[i] = allFitnesses.innerCV_sd()
        testFitnesses[i] = allFitnesses.test()
        nFeatures[i] =
            [h.num_features() for h in foldHofs[i]]
    return adjusterLearner.fit(
        originalFitnesses, stdDevs, nFeatures,
        testFitnesses, weights)
```

---

partitioned into folds. `DosaMO` then executes the `tuningOptimizer` on each fold and collects the results, that are a set of solutions for each fold. The set of solutions returned by each optimizer is optimizer-dependent in general, but in our experiments we used the Pareto front of all the solutions that were explored. For each objective `obj` the algorithm assigns weights to the solutions: for each `tuningOptimizer` result set `h`, the solutions receive a weight that is proportional to the partial derivative of the HV of the belonging result set with respect to the solution and the current objective `obj`. An adjuster regression model is trained for the current objective `obj` using the function `trainAdjuster` that receives in input a regression model `adjusterLearner`, the current objective `obj`, the `tuningOptimizer` solution set for each fold, the data including folds information (`foldsData`), and the weights of the samples (`weights`). The returned `adjuster` predicts how much the fitness of a solution changes between the training sample set and an unseen testing sample set. The function `trainAdjuster` has its own detailed description below. Using the previous objective `obj` and the `adjuster`, a new adjusted objective is created that when evaluating a solution first uses the old fitness function to compute a temporary fitness, then adjusts this fitness by subtracting the prediction of the `adjuster`. Finally, `DosaMO` runs the `mainOptimizer` on the whole `trainingData`, using the adjusted objectives instead of the original objectives.

The `trainAdjuster` function trains a fitness adjuster regression model for one of the objectives using as samples the solutions resulting from running the `tuningOptimizer` on all the folds defined by `foldsData`. The solutions are weighted: each solution has a weight proportional to the HV partial derivative with respect to the considered objective, with the weights for each fold that sum to 1.

For each fold `i` as defined by `foldsData`, `trainAdjuster` prepares the samples for training the adjuster regressor (Pseudocode 3). The samples are prepared separately for each fold, then used together in training. The function `evaluateWithCV` assigns two fitnesses to each solution: one computed on the train data of the current fold `i`, and another computed on the test data. It also computes the standard deviation of the train fitness computed with the bootstrap method. The differences between train and test performance are the values that the regression will learn to predict. In order to do that the regression has 3 input

---

**Pseudocode 4** NSGA\* class definition.

```
class Nsga* inherits MultiObjectiveOptimizer

method new(
    nGenerations, popsize, featureImportance,
    sort, tournament, mutation,
    cloneRepurposing):
    self.nGenerations = nGenerations
    self.popsize = popsize
    self.featureImportance = featureImportance
    self.sort = sort
    self.tournament = tournament
    self.mutation = mutation
    self.cloneRepurposing = cloneRepurposing

method optimize(objectives, trainingData):
    hof = createNewHallOfFame()
    featImp = self.featureImportance(trainingData)
    pop = createNewRandomIndividuals(
        featImp, self.popsize)
    for i::self.nGenerations:
        pop = evaluate(pop, objectives)
        pop = self.sort(pop)
        offspring = self.tournament(pop)
        offspring = crossover(offspring)
        offspring = self.mutation(offspring)
        pop = concat(pop, offspring)
        if self.cloneRepurposing:
            pop = ReplaceClonesWithNewIndividuals(
                pop, featImp)
        pop = evaluate(pop, objectives)
        hof = updateHallOfFame(hof, pop)
        pop = self.sort(pop)
        pop = pop[1..self.popsize]
    return hof
```

---

meta-features: the original fitness, i.e. the performance on the train data, the standard deviation of the original fitness, and the number of the features that are included in the solution (the number of genes that are included in the biomarker in our case study).

The DOSA-MO is a wrapper algorithm that uses an MO optimizer to produce the training data for the regression models and a (possibly different) MO optimizer that will run with the adjusted objectives to produce the final results for the user. In our case study we used NSGA3-CHS algorithm as both tuning and main optimizer. The parameters of population size and number of generations were lower for the tuning optimizer in order to reduce the computational load. Modified Non-Dominated Sorting Genetic Algorithm II (NSGA2) algorithms NSGA2-CH and NSGA2-CHS were first introduced in [5] and further validated in [6]. In this work we analogously modified NSGA3 with the same capabilities of NSGA2-CHS obtaining NSGA3-CHS. For this purpose we designed a more general algorithm: NSGA\*.

Pseudocode 4 describes the algorithm NSGA\*, a further generalization of the generalized Nsga2 algorithm presented in [5]. NSGA\* generalizes NSGA2 and NSGA3, together with their modifications NSGA2-CH, NSGA2-CHS, NSGA3-CH, and NSGA3-CHS. The differences between the algorithms derived from NSGA2 and NSGA3 reside in the tournament and in the sort routines. NSGA2 uses a tournament by position: the individuals are paired if their position divided by 2 is the same [16]. NSGA3 uses a random tournament where pairs are selected ran-

domly with replacement [15]. Both NSGA2 and NSGA3 use a hierarchical sort where individuals are first sorted by their domination front. The secondary sorting is different: it is based on the crowding distance in NSGA2 [16] and on reference point niching in NSGA3 [15]. The modifications CH and CHS use a primary sorting according to the clone index ([5]), and the original sorting of NSGA2 or NSGA3 as secondary sorting.

The class `Nsga*` has a constructor that memorizes the algorithm-specific parameters and strategies: the number of generations, population size, feature importance function, sorting algorithm, tournament strategy, mutation operator, and flag for the use of clone repurposing. Since the DOSA-MO uses adjusted objective functions, NSGA\* takes an `objectives` object in input, it is used by the `evaluate` function and represents the NSGA\* ability to run with different objective functions. The algorithm is similar to generalized NSGA2 [5], but with the added possibility of personalizing the tournament strategy, sorting algorithm and objectives. It can be used as inner MO optimizer by the DOSA-MO and can be specialized also as NSGA3, NSGA3-CH, and NSGA3-CHS.

## 2.2.2 Considered adjusting regression models

Five regression models have been considered in the experimental validation for adjusting the fitness functions. All of them use the sample weights, computed by the DOSA-MO as the partial derivatives of the HV as explained above.

**Dummy.** The simplest regression model learns the weighted median.

**ptree.** The pruned decision tree regression model, minimizing the weighted absolute error. The tree is pruned with the Minimal Cost-Complexity Pruning algorithm [17]. The complexity parameter for the pruning is optimized by running a 5-fold CV on its training data.

**RFReg.** The random forest regression, minimizing the weighted absolute error.

**SVR.** The epsilon-support vector regression with Gaussian kernel type [18]. It uses an l2 regularization penalty. We use the default parameters of  $C = 1$  and  $\epsilon = 0.1$ .

**rSVR.** The randomized SVR uses random search with 5-fold CV on its training data to optimize the parameters  $C$  and  $\epsilon$ .

We include in the tests the unadjusted NSGA3-CHS, that is equivalent to use regression models that predict always 0 (shortened as “zero” in the following).

## 2.2.3 Limit the computational overhead for adjustments

Running the tuning optimizer in a nested k-fold CV in order to generate the samples used for training the regression models imposes a computational overhead with

respect to the cost of running the main optimizer without any adjustment to the fitness functions. An high overhead could make the DOSA-MO impractical in real cases. In our experimental validation, that uses GAs as inner MO optimizers, we limited the computational overhead by using in the tuning optimizer a smaller population and less generations than in the main optimizer.

The simplifying assumption is made for the computational cost of the GAs to be proportional to the number of individual evaluations multiplied by the number of samples used for evaluating the individuals (in training algorithms the cost is typically at least linear in the number of samples, since the algorithm has to at least iterate through them), and for the number of evaluations to be in turn proportional to the population size multiplied by the number of generations (this is in fact an upper bound considering that individuals equal to previously evaluated ones might not need to be evaluated again). We define a parameter  $\mu$  as the ratio between the number of individuals evaluated during the tuning phase with respect to the main optimization phase.

In order to have a computation time of the DOSA-MO approximately double than the time required by the unadjusted optimizer,  $\mu$  is set to 1 in our experiments.

In the case of internal-external CV, let  $\eta$  be the number of external folds. We use the same number of folds also for the generation of the training samples for the regression models, so the tuning optimizer is executed that number of times for each execution of the DOSA-MO. Let  $\rho$  be the population size for the main optimizer, and  $\rho'$  the population size for the tuning optimizer. We compute  $\rho'$  with the following equation.

$$\rho' = \left\lfloor \rho \sqrt{\frac{\mu}{\eta - 1}} \right\rfloor \quad (3)$$

Where  $\lfloor \cdot \rfloor$  is the round to the nearest integer operation.

Let  $\gamma$  be the number of generations for the main optimizer, and  $\gamma'$  the number of generations for the tuning optimizer. Similarly, we compute  $\gamma'$  with the following equation.

$$\gamma' = \left\lfloor \gamma \sqrt{\frac{\mu}{\eta - 1}} \right\rfloor \quad (4)$$

According to the previous assumptions, the cost of running the main optimizer  $c$  is  $c = \rho\gamma m$ , with  $m$  being the number of samples. The cost of running all the iterations of the tuning optimizer  $c'$  is the following.

$$c' = \eta\rho'\gamma' \frac{m(\eta - 1)}{\eta} = \left\lfloor \rho \sqrt{\frac{\mu}{\eta - 1}} \right\rfloor \left\lfloor \gamma \sqrt{\frac{\mu}{\eta - 1}} \right\rfloor m(\eta - 1) \quad (5)$$

Ignoring the round operations that have just a small contribution it is easy to verify the desired ratio.

$$c' \approx \rho\gamma\mu m = mc \quad (6)$$

So the computational cost of using the adjusted optimizer wrapper, including the main optimizer, is approximately equal to  $(1 + m)c$ , or even lower if the cost of

evaluating individuals grows more than linearly in the number of samples.

The external validation is faster since it does not have an outer k-fold CV. For the external validations, we have arbitrarily set  $\eta = 5$ , resulting in samples for regression training to be acquired from 5 folds.

## 2.2.4 Computing the fitness functions variance

One of the high-level features of the solutions, used for prediction by the adjustment regression models, is the standard deviation of the original fitness.

Collecting the fitnesses of an individual (feature set in our case study) on the different folds and computing the standard deviation of a performance metric (for example, the balanced accuracy or the concordance index), would have a limited precision because of the number of the folds, that is necessarily contained to have an acceptable computational time. On the other hand, increasing the fold count would heighten computational demands, as the inner models need to be trained for each fold. Additionally, each fold fitnesses would be estimated on a smaller left-out sample set. Moreover, performing repeated CV would increase the number of evaluations but the same test samples would be reused.

Our method to calculate each solution's fitness variances addresses these issues by performing bootstrap analysis within each fold of the k-fold CV, followed by aggregation of the results across all folds. The following steps describe how one of the fitnesses for one of the individuals is computed.

1. For each of the folds there are training and testing samples. The inner model is optimized on the training data. The fitness is computed on the whole testing data, then by using bootstrap on the testing data, the variance for the fitness is computed.
2. The results from each of the folds are aggregated considering the folds as strata in a stratified bootstrap. The variances in the different folds are assumed uncorrelated and combined with the equation for the variance of the mean of uncorrelated random variables: the variance of the mean is the sum of the variances divided by the square of the number of folds.

When cross-validating, there are two sources of performance variance: the composition of the training set affects the training process thus can lead to different predictors, and indirectly to different expected performance, while the composition of the test set directly impacts the expected performance [19]. It is known that there is no unbiased estimator of the variance of k-fold CV [20]. The described procedure accounts for the variance explained by the limited number of test samples, but only partly for the variance explained by the limited number of training samples, since the training samples are reused in different folds (this is a well known unavoidable limitation in k-fold CV [21]) and the bootstrap is applied on the test sets but not on the training sets, to avoid incurring in unfeasible computational costs. Despite the limitations, this

estimation of the fitness variance appeared predictive of the overestimation in our experiments, thus justifying its inclusion in the high level features for the overestimation prediction.

### 2.3 Benchmarking datasets: description and pre-processing

We conducted benchmark studies with two primary goals: firstly, to identify gene expression-based biomarker sets for classifying cancer subtypes, and secondly, to determine similar biomarker sets for survival prediction in kidney and breast cancer patients.

Both case studies were addressed by using internal-external CV on TCGA [13] data. Additionally, the breast cancer case study includes an external validation with training on TCGA data and testing on SCAN-B data [14]. The latter test is particularly important as it demonstrates that the proposed algorithm can also be utilized by training predictive models for both classification and overestimation within a specific cohort (e.g. TCGA) and applying them in external cohorts (e.g. SCAN-B).

TCGA transcriptomic datasets were downloaded with the curatedTCGAData R-package version 2.0.1 from assays of type RNASeq2GeneNorm [22]. The retrieved data consists of upper-quartile-normalized TPM values. They were log transformed by applying  $\log_2(x + 1)$ .

The external gene expression-based transcriptomic dataset for breast cancer was obtained from the Gene Expression Omnibus (GEO) database (GSE96058) collected from the SCAN-B consortium. It includes FPKM log-transformed gene expression profiles. This data was already log-transformed, and we did not apply our own log-transformation to it. We applied for each value  $x$  the function  $\max(x, 0)$ , then we excluded the genes with less than 30% non-zero values.

For our study, we utilized TCGA datasets specific to cancer types. For the breast cancer case study, we used the TCGA-BRCA dataset. Additionally, for kidney cancer, we considered a compendium of TCGA datasets, which includes: KICH (Kidney Chromophobe), KIRC (Kidney Renal Clear Cell Carcinoma), and KIRP (Kidney Renal Papillary Cell Carcinoma).

Each transcriptomic dataset was filtered by removing genes with zero variance and gene expression values were standardized before use. For both TCGA and SCAN-B, cancer patient samples were categorized based on the PAM50 cancer subtype signature, which is used to determine specific molecular subtypes of breast cancer. The subtypes identified by PAM50 that were used for our experiments are: Luminal A (LumA), Luminal B (LumB), HER2-enriched (Her2), Basal-like (Basal, which is often referred to as triple-negative) and normal-like (Normal) cancer. For the task of classifying cancer subtypes in TCGA kidney cancer patients, we focused on distinguishing clear-cell renal cell carcinoma (ccRCC), chromophobe renal cell carcinoma (ChRCC), and papillary renal cell carcinoma, which was further divided into two subtypes based on recent studies identifying distinct clinical categories. These are referred to as PRCC T1 and PRCC T2. Additionally, we included samples of nor-

mal non-cancerous tissues. This classification system is based on a study published by Ricketts et al. [23]. Overall survival data for TCGA kidney cancer is from Liu et al. [24].

Supplementary Fig. S1 shows the first two principal components (after standardization) and the number of mRNA-based profiles for each kidney cancer subtype. Supplementary Fig. S2 shows the same information for the samples that are used when the prediction of the overall survival is an objective. In this cases the normal tissues are not used because in TCGA they are collected from healthy areas of cancer patients. Additionally, two samples are excluded because they are missing the survival information. Supplementary Fig. S3 shows the distribution of the survival outcomes in time.

Supplementary Fig. S4 shows the first two principal components (after standardization) and the number of mRNA-based profiles for each breast cancer subtype in TCGA data. The same information for the SCAN-B dataset is shown in Supplementary Fig. S5.

### 2.4 Experimental setup

Our tests include three datasets: TCGA kidney used for internal-external CV, TCGA breast used for internal-external CV and for the training part of external validation, and SCAN-B used for the testing part of external validation. The datasets and their preprocessing are described in Section 2.3.

Our MO problems include three objectives that measure the ability to classify the cancer subtypes, the ability to predict the overall survival, and the feature set size. These objectives are measured in the  $[0, 1]$  interval as per assumption of the algorithms. Subtype classification is measured with the balanced accuracy, survival prediction with the c-index, while the feature set size with the novel metric root-leanness. It was measured with leanness in a previous study [5]. The leanness is a value between 0 and 1 with higher values for solutions with less features. It is defined as  $1/(n + 1)$ , with  $n$  being the number of features. The leanness decreases sharply as the feature set size increases; consequently, the CHV is strongly impacted by the accuracy of the smallest biomarkers that use 1 or 2 features. Taking this into account, Cattelani et al. [6] provided an additional evaluation of the same solution sets by using a different measure for the impact of the number of features:  $1/(\sqrt{n} + 1)$ . This soft-leanness decreases less sharply as the set size increases. the root-leanness is even smoother than the soft-leanness and is defined as the root of the leanness:  $\sqrt{1/(n + 1)}$ .

The program performs a 3-fold CV repeated 3 times for internal-external CV on the kidney dataset, while, for uniformity with previous works [5, 6], a 5-fold CV on the breast dataset. The MO optimizer used is the DOSA-MO, wrapping NSGA3-CHS as both tuning optimizer and main optimizer (Section 2.2.1). The hyperparameters of NSGA3-CHS used as main algorithm are population 500, generations 500, and 3 folds used for the evaluation of the individuals. The hyperparameters related to the tuning phase of the DOSA-MO are set as explained in Section 2.2.3. Each test is repeated 6 times,

one for each of the 5 regressors described in Section 2.2.2, and one for the zero regressor (equivalent to the unadjusted NSGA3-CHS).

In this study, for the task of cancer subtype classification, NB and SVM were selected as the inner models. Meanwhile, Cox was employed as the inner model for the task of cancer survival prediction. The SVM uses the Radial basis function kernel, balanced class weight, and l2 regularization parameter  $C = 1$ .

The considered experimental setups are listed in Table 1.

In each k-fold CV execution, samples are stratified based on the included objectives. For balanced accuracy, stratification is by cancer subtype, and for overall survival, by survival status and time (binned into high or low for evenly distributed death events). If both these objectives are present, stratification combines both criteria.

During each GA call, for classification objectives, features not passing an ANOVA test ( $p$ -value 0.05) are removed from the current training samples. For survival objectives, features that fail a Wald test ( $p$ -value 0.05) in a Cox regression are discarded. If both objectives apply, features failing both tests are dropped.

For each combination of validation type, datasets, objectives, adjusting regression model, and classification inner model, we calculated the MO performance error and Pareto delta to assess the difference between expected and actual test sample performance. The CHV was also computed as an overall performance indicator for the solution sets. To the best of our knowledge it is the only proposed generalization of both HV to CV scenarios and of single-objective CV to MO [6]. CHV takes into account the differences between the performance expected by the optimization algorithm and measured on the test samples and preserves the HV appreciated properties, in particular the strict monotonicity [25]: if a set of solutions is strictly better, the CHV is guaranteed to be higher.

## 3 Results

We compare 6 different instantiations of the DOSA-MO by varying the regression model used for the adjustment, including the zero regression model, equivalent to not applying any adjustment. The regression models receive as input 3 high level features of the solutions that are correlated with the overestimation: the original fitness, its standard deviation, and the cardinality of the feature set (number of genes). For an example of correlations between the high level features and the overestimation see Supplementary Section 2. We use NSGA3-CHS as wrapped model. The tests are repeated with 8 different combinations of validation type, datasets, objectives, and classification inner model. For each combination we report two measures for the accuracy of the fitness estimation: the MO performance error (Section 3.1.1) and the Pareto delta (Section 3.1.2). Additionally, we report a measure of overall performance of the optimizers: the CHV (Section 3.2). In Section 3.3 we show in detail the

performance of the best solution sets from the external validation.

### 3.1 Overestimation in feature selection for biomarker discovery

We measured the error of the performance estimates with two methods: the MO performance error and the Pareto delta, both of which are detailed in the methods section. In summary, the MO performance error quantifies the discrepancy between the HV computed on the train performance and the CHV. On the other hand, the Pareto delta is a metric that captures differences between expected performance and actual performance on new samples, focusing on the more granular level of individual solutions.

#### 3.1.1 Comparative analysis using multi-objective performance error

The MO performance error evaluates the DOSA-MO algorithm’s capability in reducing the performance estimation error. The MO performance error quantifies the discrepancy between two key metrics that each summarize a solution set’s fitness in a single value: the HV computed from training performance and the CHV (see Section 2.1.1). We report this error across various experimental setups. The MO performance error outcomes from various regression-based adjusters are reported in Fig. 1.

Significantly, ptree and RFReg often outperform other overestimation predictors in terms of MO performance error. Therefore, for users prioritizing the reduction of MO performance error, ptree and RFReg regressors emerge as particularly effective choices. In comparison, SVR-based estimators of overfitting appear to be the least effective when considering MO performance error. Moreover, it is interesting to note that ptree and RFReg regressors perform particularly well in multi-cohort biomarker discovery and validation scenarios. In these scenarios, an external dataset (SCAN-B) is used to evaluate the models (biomarker sets), with their performance adjusted based on what has been learned from the initial TCGA-BRCA cohort.

#### 3.1.2 Comparative analysis using the Pareto delta

The Pareto delta measure (Section 2.1.2) considers the difference between a model’s predicted and actual test performance for each individual solution. This makes it particularly relevant and more informative than the MO performance error in the common situation when users are presented with multiple solutions but will ultimately select only one. A comparison of the Pareto delta measure across various experimental setups is detailed in Fig. 2.

In the DOSA-MO framework, ptree and RFReg regressors consistently outshine others when reducing performance estimation error, aggregated on a solution by solution basis with the Pareto delta. Interestingly, while no significant differences are noted in MO performance

Datasets	Objectives	Inner models	Validation
Kidney cancer	Subtype classification (balanced accuracy), set size (root-leanness)	NB	CV
Kidney cancer	Overall survival (c-index), set size (root-leanness)	Cox	CV
Kidney cancer	Subtype classification (balanced accuracy), overall survival (c-index), set size (root-leanness)	NB/COX	CV
Breast cancer	Subtype classification (balanced accuracy), set size (root-leanness)	NB, SVM	CV
Breast cancer	Subtype classification (balanced accuracy), set size (root-leanness)	NB, SVM	External (SCAN-B)
Breast cancer	Overall survival (c-index), set size (root-leanness)	Cox	CV

Table 1: Overview of datasets, objectives, inner models, and validation methods used in the analysis of kidney and breast cancer.

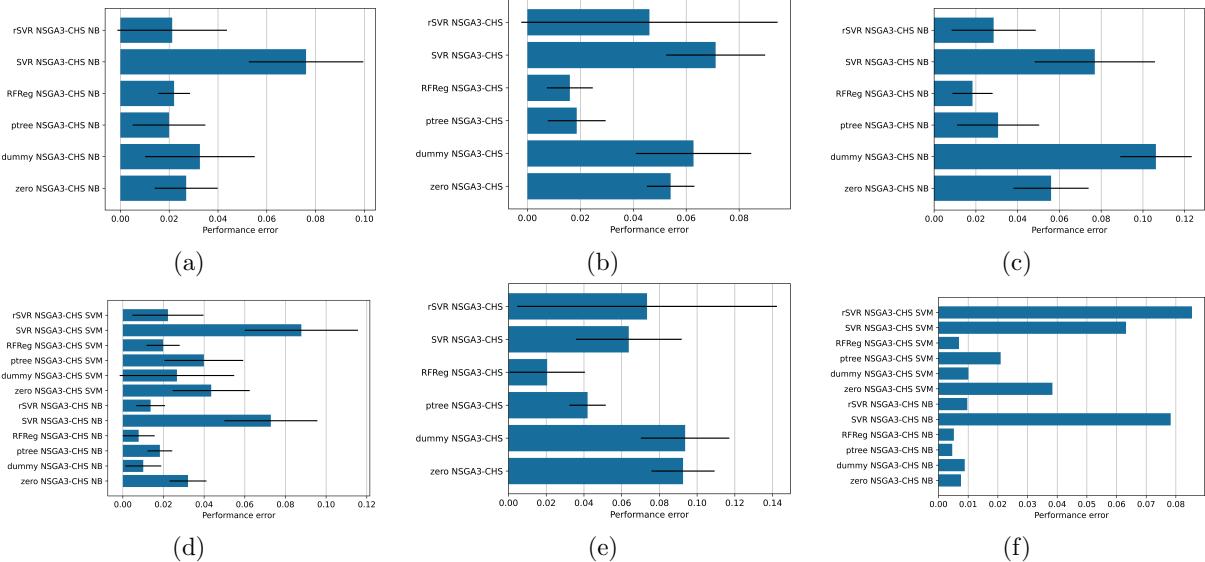


Figure 1: MO performance error results for internal-external CV (a-e) and external validation (f). (a) Kidney cancer, subtype classification and root-leanness. (b) Kidney cancer, overall survival prediction and root-leanness. (c) Kidney cancer, overall survival prediction, subtype classification, and root-leanness. (d) Breast cancer, subtype classification and root-leanness. (e) Breast cancer, overall survival prediction and root-leanness. (f) External validation for breast cancer, subtype classification and root-leanness. Error bars represent standard deviation between folds.

error between the dummy model and the unadjusted optimizer, the Pareto delta metric reveals that all kidney cancer setups with proper regressors have a lower Pareto delta compared to the unadjusted optimizer. rSVR performs better than the zero regressor in all scenarios except one. RFReg and ptree always yield lower Pareto deltas than the zero regressor, including in breast cancer setups, highlighting their promise in both overestimation measures.

### 3.2 Evaluating DOSA-MO’s effectiveness in MO feature selection

The previous results highlight DOSA-MO’s success in reducing performance estimation errors in ML-driven feature selection for biomarker discovery. This section evaluates if DOSA-MO’s dual-stage framework also enhances the quality of the produced feature set, that is the fi-

nal output for biomarker selection models considered for clinical validation. To assess DOSA-MO’s impact on the overall feature selection process, we calculated the CHV for each experimental setup (see Fig. 3).

Tree-based regressors effectively reduce estimation errors but do not significantly improve the quality of solution sets generated by the MO optimizer. This indicates that using these regressors for overestimation prediction may not always enhance solution sets, despite reducing quality estimation errors. Interestingly, dummy and SVR-based MO optimizers consistently outperform the CHV of unadjusted models, highlighting the effectiveness of overestimation prediction in DOSA-MO for superior feature selection outcomes. Specifically, SVR outperforms the zero model in all kidney setups and most breast setups, except in scenarios involving the overall survival objective. Notably, SVR yields optimal results in setups excluding survival prediction, while the dummy

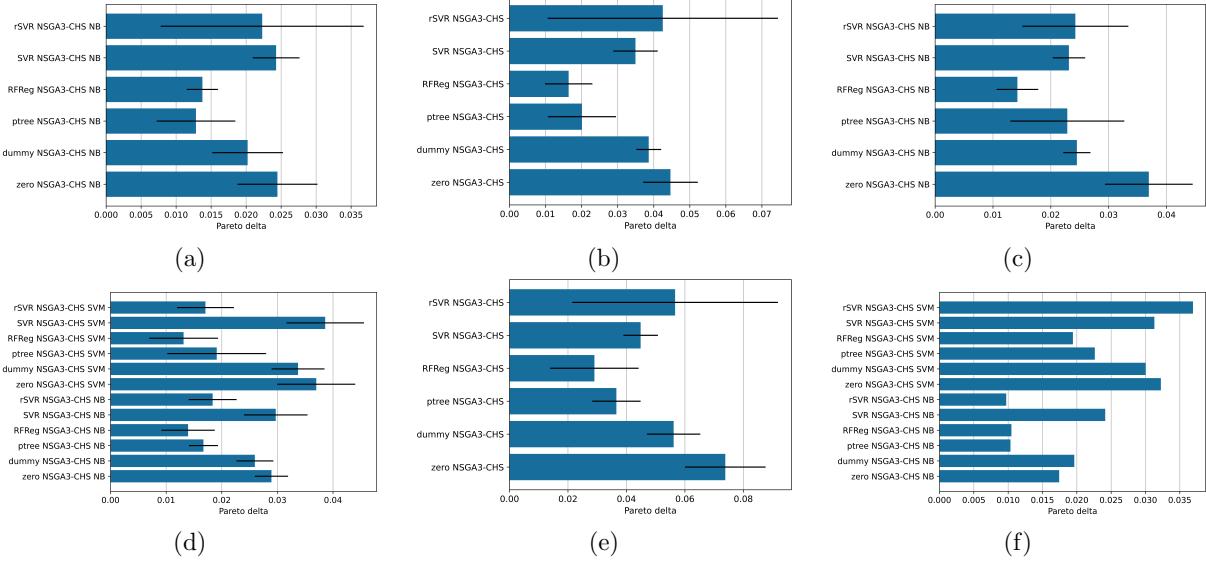


Figure 2: Pareto delta results for internal-external CV (a-e) and external validation (f). (a) Kidney cancer, subtype classification and root-leanness. (b) Kidney cancer, overall survival prediction and root-leanness. (c) Kidney cancer, overall survival prediction, subtype classification and root-leanness. (d) Breast cancer, subtype classification and root-leanness. (e) Breast cancer, overall survival prediction and root-leanness. (f) External validation for breast cancer, subtype classification and root-leanness. Error bars represent standard deviation between folds.

model excels in setups that include survival prediction.

### 3.3 Advancing biomarker discovery using multi-cohort studies.

In this section, we examine a case study that utilizes two distinct cohorts, TCGA-BRCA and SCAN-B, for external validation in biomarker discovery. The approach focuses on learning models and overestimation during biomarker identification within one cohort (like TCGA) and then applying this knowledge to another cohort (such as SCAN-B). This shows the reliability of biomarker discovery across different cohorts.

As previously shown, in the external validation for classification of breast cancer subtypes the DOSA-MO with SVR as fitness adjustment regression model and SVM as inner classifier provides the best overall performance measured with the CHV, while, keeping the SVM as inner classifier, using the RFReg it is possible to obtain the more accurate prediction of the performance on a new cohort, with respect to both MO performance error and Pareto delta. The effect of the adjustment can be appreciated by looking at the solution sets presented in Fig. 4. The balanced accuracy expected by the optimizer ("internal cv" in the figures) is monotonically increasing with the number of features because the solutions returned by the algorithm are non-dominated. The estimation error, i.e. the vertical distance between the balanced accuracy expected by the optimizer using only the training samples and the balanced accuracy measured on the external dataset, is lower when using the DOSA-MO.

## 4 Discussion

We presented a new wrapper MO optimizer capable of adjusting the performance measures for the overestimation that is due to multiple model validations, improving the solution set that is produced. It has been validated on multiple transcriptomics configurations related to kidney and breast cancer, including external validation.

Since to the best of our knowledge there was no measure of error in the performance estimation for MO solution sets, we proposed two such measures. According to both of them, the DOSA-MO provided improved performance estimates in all the considered scenarios when a regression model based on decision trees was used to predict the overestimation.

Even the dummy regression model that learns just the weighted average of the overestimation, when used for adjusting the fitness, allows to obtain a better overall performance than the unadjusted optimizer in 7 out of 8 experimental setups, as measured by the CHV. The SVR regression model also provides for a better CHV than the unadjusted optimizer in 7 out of 8 setups. The dummy regression model is the best in the 3 setups that include survival prediction, while the SVR is the best in the 5 setups that have subtype classification and economy of features as objectives.

The rSVR regression model selects the hyperparameters for the regularization strength of the SVR using random search. It performed poorly in our tests. A regularization that works well while cross-validating on solutions obtained by the same unadjusted optimizer, is not as effective when the GA runs with a bigger population, more generations, and the adjustment applied to the evaluation of the individuals. More conservative regression models, like the dummy, or the SVR with its

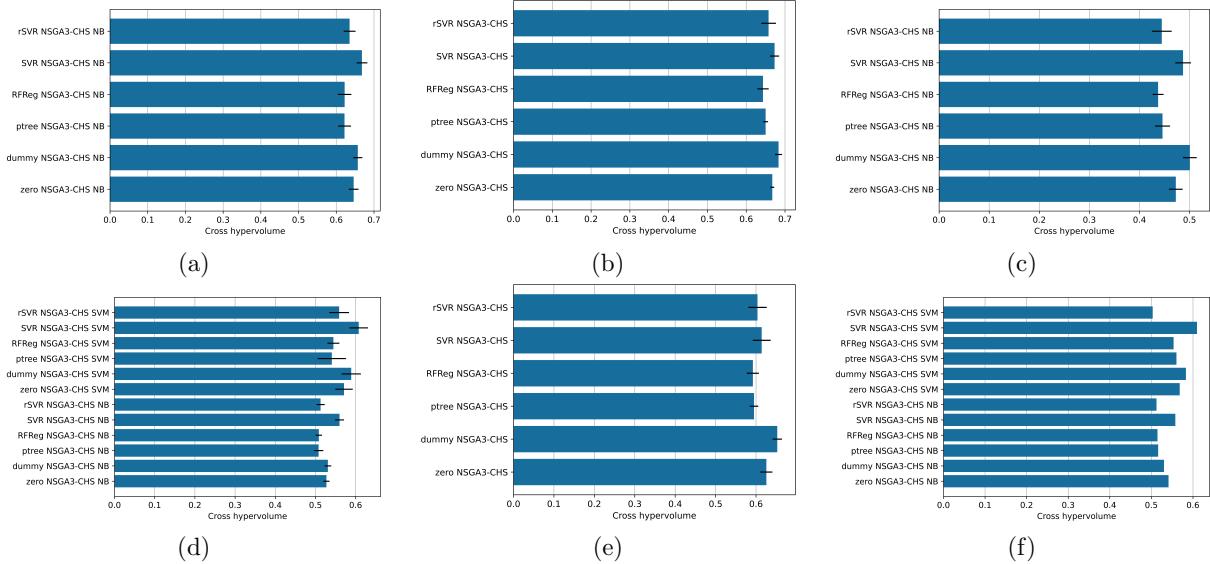


Figure 3: CHV results for internal-external CV (a-e) and external validation (f). (a) Kidney cancer, subtype classification and root-leanness. (b) Kidney cancer, overall survival prediction and root-leanness. (c) Kidney cancer, overall survival prediction, subtype classification and root-leanness. (d) Breast cancer, subtype classification and root-leanness. (e) Breast cancer, overall survival prediction and root-leanness. (f) External validation for breast cancer, subtype classification and root-leanness. Error bars represent standard deviation between folds.

strong default regularization, have proven more effective in our tests.

Predicting the overestimation is more complex when the adjustment to the fitness is applied during the optimization as in our approach. The regression models are trained on sample distributions, but then they apply the adjustment results to different solutions affecting their selection. We cannot expect for the regression models to be as precise in areas of the solution space that have not been explored when creating the training set. This applicability domain issue brings an added difficulty.

Adjusting the fitness impacts the exploration of the optimizer, and the explored solutions will have a different distribution from the one on which the regressors were trained, even in the areas included in the applicability domain. This is a moving target problem, often solved in other cases by optimizing in small steps, for example when applying back-propagation to artificial neural networks across multiple epochs. It is not practical to run many iterations of computationally expensive MO GAs, but strategies involving faster algorithms may be explored. Since the solutions can potentially still be used for training the regression models also in the following steps, techniques may be mutuated from on-line learning algorithms. More research is needed to improve how the applicability domain and moving target problems are addressed. The proposed algorithm can be applied to any data and problem and has been experimentally validated on cancer transcriptomics data for subtype classification and survival prediction. Further experimental validation is needed to assess its performance in other domains.

## References

- [1] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.
- [2] Kalyanmoy Deb and Kalyanmoy Deb. *Multi-objective Optimization*, pages 403–449. Springer US, Boston, MA, 2014. ISBN 978-1-4614-6940-7. doi:10.1007/978-1-4614-6940-7\_15. URL [https://doi.org/10.1007/978-1-4614-6940-7\\_15](https://doi.org/10.1007/978-1-4614-6940-7_15).
- [3] Angela Serra, Serli Önlü, Paola Festa, Vittorio Fortino, and Dario Greco. MaNga: a novel multi-niche multi-objective genetic algorithm for QSAR modelling. *Bioinformatics*, 36(1):145–153, 06 2019. ISSN 1367-4803. doi:10.1093/bioinformatics/btz521. URL <https://doi.org/10.1093/bioinformatics/btz521>.
- [4] V Fortino, G Scala, and D Greco. Feature set optimization in biomarker discovery from genome-scale data. *Bioinformatics*, 36(11):3393–3400, 04 2020. ISSN 1367-4803. doi:10.1093/bioinformatics/btaa144. URL <https://doi.org/10.1093/bioinformatics/btaa144>.
- [5] Luca Cattelani and Vittorio Fortino. Improved nsga-ii algorithms for multi-objective biomarker discovery. *Bioinformatics*, 38(Supplement\_2):ii20–ii26, 09 2022. ISSN 1367-4803. doi:10.1093/bioinformatics/btac463. URL <https://doi.org/10.1093/bioinformatics/btac463>.
- [6] Luca Cattelani, Arindam Ghosh, Teemu Rintala, and Vittorio Fortino. Improving biomarker selection for cancer subtype classification through multi-objective optimization. 10 2023. doi:10.36227/techrxiv.24321154.v2. URL <https://doi.org/10.36227/techrxiv.24321154>.

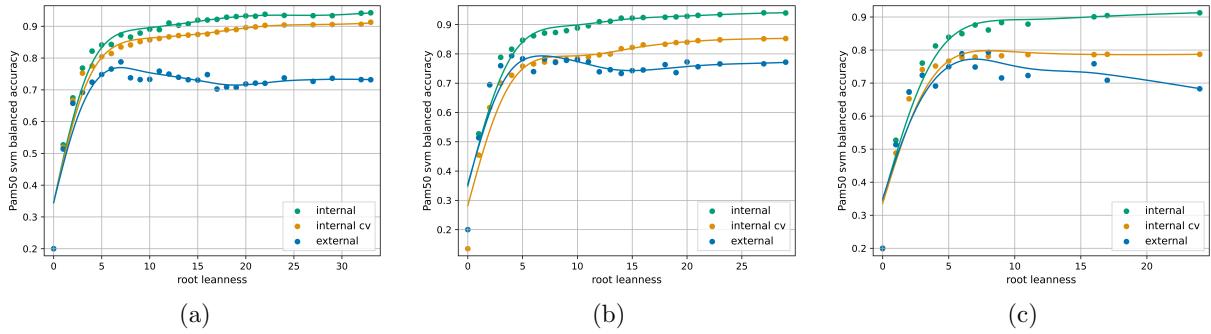


Figure 4: Scatter plots depicting solutions from external validation on breast cancer transcriptomics data using SVM as inner model. MO optimization of balanced accuracy for subtypes prediction and root-leanness. Horizontally, the number of features is depicted for simplicity. For each solution it is shown the performance measured in the inner CV, i.e. the performance expected by the optimizer, the performance of the model trained on the TCGA breast set and tested on the same set, and the performance of the same model on the external SCAN-B set. The lines are interpolating splines. (a) Using the unadjusted optimizer. (b) Using SVR as regression model for fitness adjustment. (c) Using RFReg as regression model for fitness adjustment.

- [7] Damjan Krstajic, Ljubomir J Buturovic, David E Leahy, and Simon Thomas. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6(1):1–15, 2014.
- [8] Tzu-Tsung Wong. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9):2839–2846, 2015. ISSN 0031-3203. doi:<https://doi.org/10.1016/j.patcog.2015.03.009>. URL <https://www.sciencedirect.com/science/article/pii/S0031320315000989>.
- [9] Tzu-Tsung Wong and Po-Yang Yeh. Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1586–1594, 2020. doi:10.1109/TKDE.2019.2912815.
- [10] Ioannis Tsamardinos, Amin Rakhshani, and Vincenzo Lagani. Performance-estimation properties of cross-validation-based protocols with simultaneous hyperparameter optimization. *International Journal on Artificial Intelligence Tools*, 24(05):1540023, 2015.
- [11] Ryan J. Tibshirani and Robert Tibshirani. A bias correction for the minimum error rate in cross-validation. *The Annals of Applied Statistics*, 3(2):822 – 829, 2009. doi:10.1214/08-AOAS224. URL <https://doi.org/10.1214/08-AOAS224>.
- [12] Ioannis Tsamardinos, Paulos Charonyktakis, Georgios Papoutsoglou, Giorgos Borboudakis, Kleanthi Lakiotaki, Jean Claude Zenklusen, Hartmut Juhl, Ekaterini Chatzaki, and Vincenzo Lagani. Just add data: automated predictive modeling for knowledge discovery and feature selection. *NPJ precision oncology*, 6(1):38, 2022.
- [13] Carolyn Hutter and Jean Claude Zenklusen. The cancer genome atlas: creating lasting value beyond its data. *Cell*, 173(2):283–285, 2018.
- [14] Christian Bruegger, Johan Vallon-Christersson, Dorthe Grabau†, Anna Ehinger, Jari Häkkinen, Cecilia Hegardt, Janne Malina, Yilun Chen, Pär-Ola Bendahl, Jonas Manjer, Martin Malmberg, Christer Larsson, Niklas Loman, Lisa Rydén, Åke Borg, and Lao H. Saal. Clinical value of rna sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter sweden cancerome analysis network—breast initiative. *JCO precision oncology*, 2:1–18, Mar 2018. doi:10.1200/PO.17.00135. URL <https://doi.org/10.1200/PO.17.00135>. PMID: 32913985.
- [15] Kalyanmoy Deb and Himanshu Jain. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: Solving problems with box constraints. *IEEE Transactions on Evolutionary Computation*, 18(4):577–601, 2014. doi:10.1109/TEVC.2013.2281535.
- [16] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6 (2):182–197, 2002. doi:10.1109/4235.996017.
- [17] L Breiman, JH Friedman, R Olshen, and CJ Stone. Classification and regression trees. 1984.
- [18] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. In M.C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.
- [19] Juan D. Rodriguez, Aritz Perez, and Jose A. Lozano. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):569–575, 2010. doi:10.1109/TPAMI.2009.187.
- [20] Yoshua Bengio and Yves Grandvalet. Bias in estimating the variance of k-fold cross-validation. In *Statistical modeling and analysis for complex data problems*, pages 75–95. Springer, 2005.
- [21] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(none):40 – 79, 2010. doi:10.1214/09-SS054. URL <https://doi.org/10.1214/09-SS054>.
- [22] Marcel Ramos, Ludwig Geistlinger, Sehyun Oh, Lucas Schiffer, Rimsha Azhar, Hanish Kodali, Ino de Brujin, Jianjiang Gao, Vincent J Carey, Martin Morgan, et al. Multiomic integration of public oncology databases in bioconductor. *JCO Clinical Cancer Informatics*, 1:958–971, 2020.

- [23] Christopher J Ricketts, Aguirre A De Cubas, Huihui Fan, Christof C Smith, Martin Lang, Ed Reznik, Reanne Bowlby, Ewan A Gibb, Rehan Akbani, Rameen Beroukhim, et al. The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell reports*, 23(1):313–326, 2018.
- [24] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated tega pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018.
- [25] Miqing Li and Xin Yao. Quality evaluation of solution sets in multiobjective optimisation: A survey. *ACM Comput. Surv.*, 52(2), mar 2019. ISSN 0360-0300. doi:10.1145/3300148. URL <https://doi.org/10.1145/3300148>.

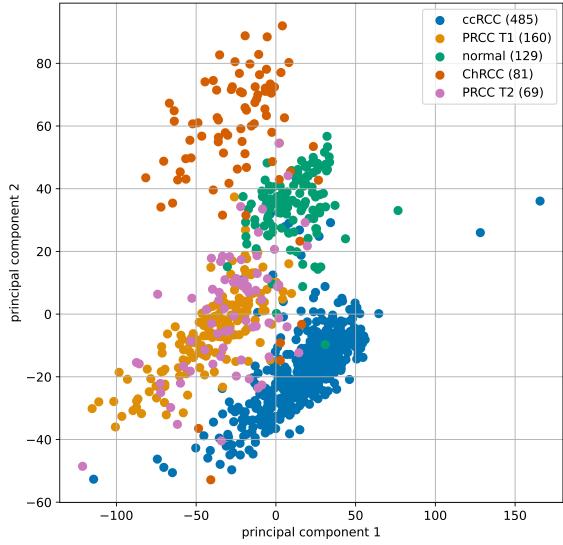
# DOSA-MO: Dual-stage Optimizer for Systematic overestimation Adjustment in Multi-Objective problems improves biomarker discovery, supplementary material

Luca Cattelani<sup>\*1</sup> and Vittorio Fortino<sup>†1</sup>

<sup>1</sup>Institute of Biomedicine, School of Medicine, University of Eastern Finland, Kuopio, 70211, Finland.

---

<sup>\*</sup>Corresponding Author: luca.cattelani@uef.fi  
<sup>†</sup>Corresponding Author: vittorio.fortino@uef.fi



Supplementary Figure S1: First two principal components with subtypes of all the TCGA kidney samples.

## 1 Datasets

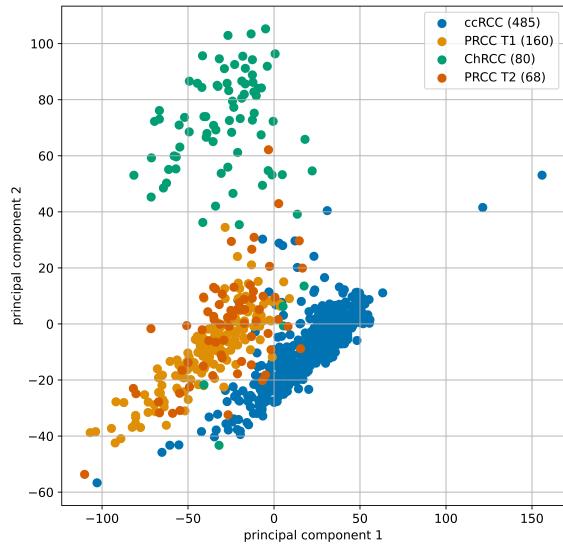
The TCGA kidney dataset at the end of curation includes 5 classes and a total of 924 samples (Supplementary Fig. S1). When the prediction of survival was included in the objectives, a smaller dataset of 793 samples was used in order to avoid the unlabelled samples (Supplementary Fig. S2). This led to the removal of all the samples of normal (healthy) tissue. Supplementary Fig. S3 shows the distribution of the survival outcomes.

The TCGA breast dataset at the end of curation includes 5 classes and a total of 1081 samples (Supplementary Fig. S4).

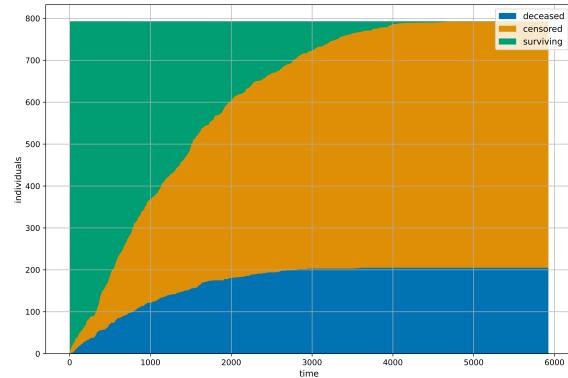
The SCAN-B dataset at the end of curation includes 5 classes and a total of 2969 samples (Supplementary Fig. S5).

## 2 High level features correlation

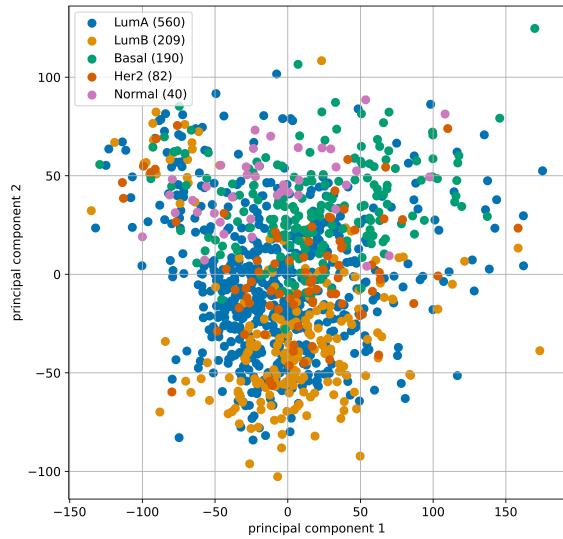
Supplementary Fig. S6 shows an example of solution set performance after executing optimization with NSGA3-CHS for kidney cancer subtypes biomarker discovery, without any performance adjustment. The optimizer explored the best trade-offs between balanced accuracy and feature set size, and used 3-fold CV internally to evaluate the solutions. The overestimation (difference between the yellow and blue dots) of the balanced accuracy increases as the number of features increases. Analogously, the overestimation increases when the estimation increases. Also the variance of the estimation, computed by bootstrapping the samples, is correlated with the overestimation: the 9 p-values of the Kendall correlation for a 3-fold CV repeated 3 times are 0.008, 0.000, 0.068, 0.020, 0.006, 0.012, 0.003, 0.034, and 0.021.



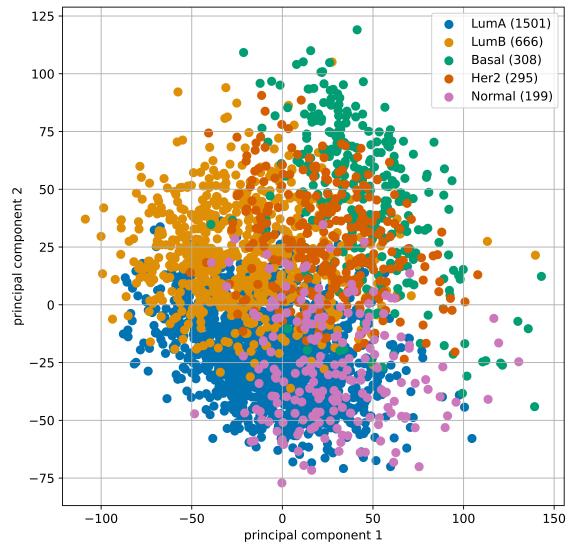
Supplementary Figure S2: First two principal components with subtypes of the TCGA kidney samples with survival labels.



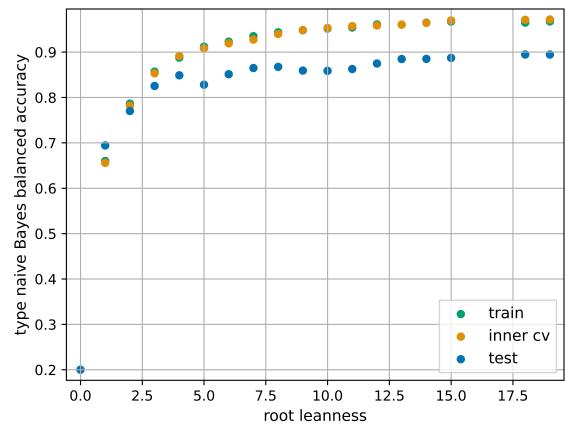
Supplementary Figure S3: TCGA kidney survival outcomes.



Supplementary Figure S4: First two principal components with subtypes of the TCGA breast samples.



Supplementary Figure S5: SCAN-B PCA.



Supplementary Figure S6: Example of scatter plot depicting solutions from CV on kidney cancer transcriptomics data. MO optimization of balanced accuracy for subtypes prediction and root-leanness. Horizontally, the number of features is shown for simplicity. For each solution it is shown the performance measured in the inner CV, i.e. the performance expected by the optimizer, the performance of the model trained on the whole training set and tested on the same set, and the performance of the same model on the testing set. The number of features, the expected performance and the overestimation are correlated.