Journal of Biomedical Informatics 144 (2023) 104457

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Supervised Relevance-Redundancy assessments for feature selection in omics-based classification scenarios

Silvia Cascianelli [*], Arianna Galzerano, Marco Masseroli

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci, 32, Milano, 20133, Italy

ARTICLE INFO

Keywords:
Feature selection
Unbalanced classification
Clinically-relevant stratification
Relevance-redundancy strategies
Transcript isoforms

ABSTRACT

Background and objective: Many classification tasks in translational bioinformatics and genomics are characterized by the high dimensionality of potential features and unbalanced sample distribution among classes. This can affect classifier robustness and increase the risk of overfitting, curse of dimensionality and generalization leaks; furthermore and most importantly, this can prevent obtaining adequate patient stratification required for precision medicine in facing complex diseases, like cancer. Setting up a feature selection strategy able to extract only proper predictive features by removing irrelevant, redundant, and noisy ones is crucial to achieving valuable results on the desired task.
Methods: We propose a new feature selection approach, called ReRa, based on supervised Relevance-Redundancy assessments. ReRa consists of a customized step of relevance-based filtering, to identify a reduced subset of meaningful features, followed by a supervised similarity-based procedure to minimize redundancy. This latter step innovatively uses a combination of global and class-specific similarity assessments to remove redundant features while preserving those differentiated across classes, even when these classes are strongly unbalanced.
Results: We compared ReRa with several existing feature selection methods to obtain feature spaces on which performing breast cancer patient subtyping using several classifiers: we considered two use cases based on gene or transcript isoform expression. In the vast majority of the assessed scenarios, when using ReRa-selected feature spaces, the performances were significantly increased compared to simple feature filtering, LASSO regularization, or even MRmr — another Relevance-Redundancy method. The two use cases represent an insightful example of translational application, taking advantage of ReRa capabilities to investigate and enhance a clinically-relevant patient stratification task, which could be easily applied also to other cancer types and diseases.
Conclusions: ReRa approach has the potential to improve the performance of machine learning models used in an unbalanced classification scenario. Compared to another Relevance-Redundancy approach like MRmr, ReRa does not require tuning the number of preserved features, ensures efficiency and scalability over huge initial dimensionalities and allows re-evaluation of all previously selected features at each iteration of the redundancy assessment, to ultimately preserve only the most relevant and class-differentiated features.

Statement of significance

Problem or Issue: Clinically-relevant classifications can suffer from uneven class distributions and huge, unbalanced dimensionalities of features, which are often noisy and highly redundant.

What is Already Known: Feature selection is fundamental to extracting predictive features for robust classification. Yet, widely-adopted univariate filters, which are computationally efficient and scalable, struggle to remove redundancy; Relevance-Redundancy approaches inspect also feature relationships to minimize global redundancy, but cannot easily scale over huge feature sizes.

What this Paper Adds: We propose ReRa, a new efficient Relevance-Redundancy approach. ReRa uses both global and supervised class-specific similarity assessments to optimize feature selection, also considering differential behaviors that can improve classifications.

1. Introduction

1.1. Background

In translational bioinformatics and computational genomics, many problems can be addressed as machine learning classification tasks

[*] Corresponding author.
E-mail address: silvia.cascianelli@polimi.it (S. Cascianelli).

https://doi.org/10.1016/j.jbi.2023.104457
Received 28 February 2023; Received in revised form 5 June 2023; Accepted 19 July 2023
Available online 23 July 2023

based on omics data used as quantitative features, like gene expression values. Key predictive features are extracted from such feature spaces before or while training a classifier to perform the desired task. Yet, initial space dimensionality is often extremely high and unbalanced compared with the restrained number of samples available, which, in addition, are mostly unevenly distributed among the classes reflecting real case occurrences. This exposes learners to the curse of dimensionality effects and to generalization leaks. Furthermore, features are very often noisy and strongly redundant, making the training longer and a classifier less robust and more subject to overfitting risk.

In such a context, feature selection (FS) strategies play a crucial role in achieving valuable results on the desired task: they can extract from the original feature space a subset of predictive features appropriate for the task by removing irrelevant, redundant, and noisy features [1–3]. Canonical univariate filters, which individually evaluate the predictive power of each feature, are the best option to ensure computational efficiency and scalability in extracting relevant features within huge feature spaces [4,5]. Still, they do not account for feature interactions and redundancy. Relevance-Redundancy FS approaches have been instead developed as multi-step and/or multi-objective filtering methods that maximize feature relevance by discarding noisy or irrelevant features for the task and minimize redundancy based on feature similarity [6–8]. Although this analysis of feature mutual behaviors overcomes the main weakness of canonical filters, multi-objective Relevance-Redundancy approaches often struggle to scale over a huge feature size and do not consider the differential relationships that can arise among features within different classes of interest.

1.2. ReRa feature selection strategy

Here we propose a novel supervised FS method based on Relevance-Redundancy assessments, named ReRa, that aims to improve feature selection using efficient filtering and multiple similarity evaluations based on the supervised information provided by the target classes. Global similarity assessments could hide a differential behavior in a specific class, which instead could be of potential interest for a classification task, especially in the case of unbalanced class distributions. Using also class-specific similarity, our proposed ReRa method discards real redundant features preserving those differentiated within any of the considered classes. Thus, it reveals any divergence between pairs of features that might have a role in providing a better class distinction but, from a global perspective only, would be ignored.

ReRa feature selection strategy uses two consecutive steps for relevance and redundancy evaluations in classification tasks, ensuring scalability and wide applicability, since it is very generalizable for different purposes and scopes. Here, we demonstrated its reliability in the context of a significant translational application, moving from transcriptomics data analysis to clinically-relevant stratification of diseased patients in order to improve their characterization and clinical handling. To assess the usefulness of our ReRa approach, it was compared with several benchmarks, including knowledge-based and statistical filters, embedded LASSO selection, and another Relevance-Redundancy approach. Specifically, this validation was carried out through two example use cases of increasing complexity, i.e., a gene-level and a transcript isoform-level expression-based stratification of breast cancer (BRCA) patients into clinically relevant target subtypes [9,10]. These example use cases demonstrate the value and efficiency of our innovative ReRa approach, beyond highlighting its capability of considering class-specific feature relationships to improve the feature selection process for the following classification task. Such characteristics make the approach precious in the context of omics dataset analysis since the amount of involved big data, requiring efficient FS methods, together with the unbalanced input dimensionality and class distribution, can prevent obtaining adequate patient stratification: yet, this is a key requirement for precision medicine and personalized treatments to face complex diseases like cancer.

1.3. State-of-the-art for supervised feature selection and Relevance-Redundancy methods

In a supervised setting, such as the one of predictive machine learning workflows, supervised label information on training samples can be used to guide the feature selection process. Supervised feature selection strategies include all the known categories of selection methods [1–3] (filters, wrappers, embedded and hybrid methods — i.e., filter + wrapper), based on label information and using a given or no interaction with the learner. Even when working with huge feature spaces, filters ensure computational efficiency and scalability by selecting features based on their relevance for the task [4,5,11–13] independently of the predictor subsequently used (i.e., without any interaction with the learner). Conversely, the computational cost of wrapper methods [14] (like recursive backward elimination or forward selection) and of most hybrid methods is prohibitive: though they could handle the relevance-redundancy trade-off in feature selection, they require training a model for each of the spaces evaluated during their recursive search. An embedded technique such as LASSO (Least Absolute Shrinkage and Selection Operator) [15], where the feature selection is integrated and optimized into the learning process using a penalization parameter, can instead provide still valuable results in a much more efficient way than wrapper methods; as drawbacks, its selection does not use a minimal redundancy strategy, is often more suitable just for the model under consideration, and requires a longer training phase.

Therefore, filters based on statistical measures or prior knowledge are mostly preferred, despite usually evaluating one feature at a time without considering any interaction or relationship among features. To fill this gap, Relevance-Redundancy feature selection approaches were developed as multi-step and/or multi-objective methods that select only highly-relevant features while discarding redundant features based on a similarity measure [6–8]. These approaches reach the simultaneous goals of maximizing feature relevance (getting rid of irrelevant or noisy features) and minimizing feature global redundancy: relevance can be well established for individual features, whereas redundancy is typically inspected by examining feature subsets and considering their global relationships, as to overcome the main limitation of filters. While building the selected feature space, multi-step implementations can iteratively evaluate the inclusion of one feature at a time, either using a simultaneous multi-objective optimization of relevance and redundancy or following two phases: first, assess the relevance and, then, handle redundancy. One of the most used approaches is the Maximum Relevance minimum redundancy strategy (MRmr) [6], originally presented as a multi-objective algorithm for feature selection in microarray gene expression data. This iteratively selects the features at the top of a relevance ranking, as long as they contribute new and non-redundant information for the desired task. In this way, the best subset of K different features is selected based on feature relationships and redundancy and is not necessarily made of the best K features, which individually have the strongest predictive value (i.e., relevance) for the target variable. Currently, this kind of strategy has different implementations, using alternative relevance measures and correlation types to estimate relevance and redundancy, respectively (e.g., [7,8]). A common example of MRmr use in computational genomics is identifying gene signatures, including genes differentially expressed or strongly characterizing a particular phenotype (compared to a normal reference), but minimizing redundant information. Nonetheless, the MRmr approach struggles to perform such kind of investigation over big genomic data, as the ones nowadays produced by Next-Generation Sequencing experiments. Differently from the here proposed ReRa approach, the MRmr simultaneous multi-objective optimization of relevance and redundancy cannot be applied over several tens of thousands of features without requiring computational and memory resources often unaffordable.

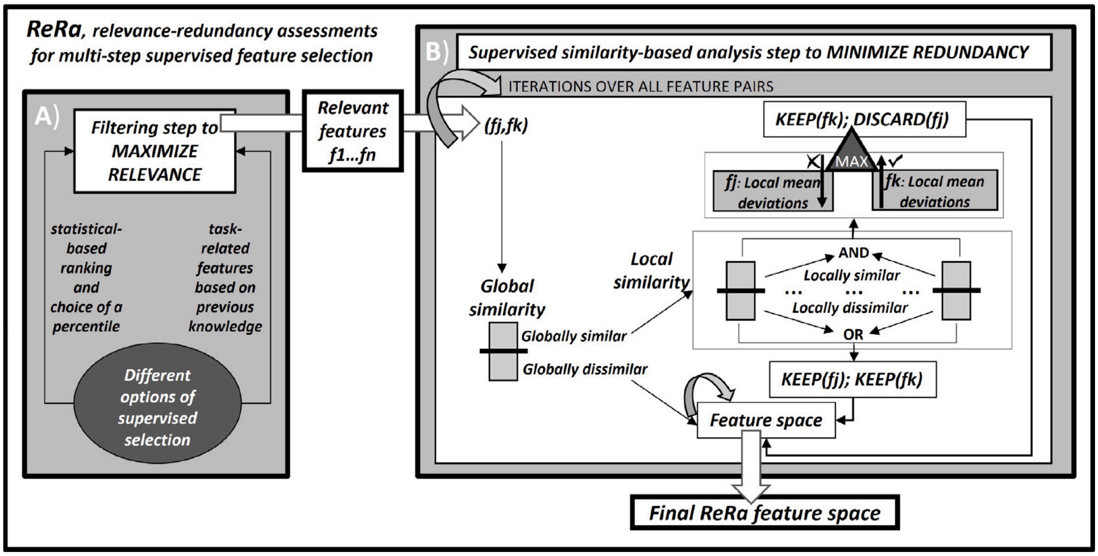
ReRa, relevance-redundancy assessments for multi-step supervised feature selection
A)
Filtering step to MAXIMIZE RELEVANCE
statistical-based ranking and choice of a percentile
task-related features based on previous knowledge
Different options of supervised selection
Relevant features f1...fn
B)
Supervised similarity-based analysis step to MINIMIZE REDUNDANCY
ITERATIONS OVER ALL FEATURE PAIRS
(fj,fk)
KEEP(fk); DISCARD(fj)
MAX
fj: Local mean deviations
fk: Local mean deviations
AND
Locally similar
Locally dissimilar
OR
Local similarity
Global similarity
Globally similar
Globally dissimilar
KEEP(fj); KEEP(fk)
Feature space
Final ReRa feature space

Fig. 1. Flowchart with the two consecutive steps of our supervised maximum relevance minimum redundancy feature selection approach, named ReRa.

2. Methods

We designed an innovative feature selection methodology named ReRa, based on supervised Relevance-Redundancy assessments, and compared it with some existing approaches of feature selection; these latter strategies are discussed in Section 2.2, whilst the next Section 2.1 is devoted to a thorough description of the here proposed ReRa approach.

2.1. Multi-step supervised feature selection with maximum relevance minimum redundancy

Our ReRa supervised feature selection approach aims to remove redundant features while preserving those relevant and differentiated across the classes of a classification scenario. In Fig. 1, we schematically illustrate the two steps of ReRa, described in detail in the following subsections.

ReRa relevance maximization is implemented as a first filtering step that ensures scalability over huge data dimensionalities. The redundancy minimization is then performed iteratively over all the relevant features and allows reevaluating and updating the selected feature space at each iteration. Similarity measures used for redundancy elimination are usually computed only globally, i.e., over all the considered samples independently of their classes; this can hide differential behavior in one or more specific classes. Instead, ReRa also considers the class labels available for supervised training to tackle feature redundancy: this is particularly useful for a classification task, especially in the case of unbalanced class distributions, where global sample assessments cannot provide sufficient indications. Furthermore, using its combination of global and class-specific similarity assessments (described in Section 2.1.2), ReRa can automatically select the features that are suitable to distinguish the classes of interest better, given whichever dimensionality of the initially redundant feature space, thus ensuring both scalability and efficiency.

2.1.1. Filtering step to maximize relevance

The first step of ReRa is a relevance-based filter, which extracts a meaningful, reduced set of features, like the ones belonging to the top N percentile of a feature ranking given by a statistical measure (see panel A of Fig. 1). Selection approaches based on relevance can be led by a supervised or unsupervised measure: in ReRa, the chosen strategy is fully supervised since known class labels used in training can drive towards a more precise selection focused on the desired classification task. Accordingly, for this first step of feature relevance maximization, we considered two alternative supervised statistical measures, i.e., Fisher Score and Mutual Information (their details are reported in Section 2.2.1). Yet, this step of relevance maximization can use other supervised strategies; any supervised filtering method can be used as ReRa first step, such as filters based on Relief strategy [16] or other statistical measures, like the Permutation of the Feature Importance [17]. In addition, a reduced set of relevant features can also be obtained according to a feature set already known to be relevant for the considered task, based on previous knowledge on the subject. Accordingly, we also explored this option (see panel A of Fig. 1), taking advantage of task-related known feature spaces for the application examples presented in the following.

2.1.2. Supervised similarity-based analysis step to minimize redundancy

Starting from an obtained relevance-based feature space, a combination of similarity analyses is performed to remove redundancy without discarding features with class-specific differentiation. Similarity assessments are both global, i.e., considering all the given samples together, and local, i.e., applied separately on each class of samples based on the supervised target labels that stratify the training data. A temporary feature space is created and updated after each global and local assessment until the end of the ReRa algorithm. The workflow of this second step of ReRa is summarized in panel B of Fig. 1.

At first, a global similarity assessment is performed by considering iteratively the pairwise correlations of all relevant features previously selected: each correlation is compared against a data-driven global threshold, whose definition is reported at the end of this subsection. If the global correlation of a feature pair is smaller than the global threshold, the two features are considered globally dissimilar and are both kept in the temporary feature space; conversely, if their similarity is equal or greater than the global threshold, they are considered globally similar, and as many local similarity assessments as the number of classes (i.e., target labels) are performed between the two features. Each of these local assessments has its own data-driven local threshold, which is compared with the local correlation of the two features. In panel B of Fig. 1, an OR operator jointly evaluates the predicates on local dissimilarity: if at least one of the local pairwise correlations is smaller than the corresponding local threshold, both features are included in the temporary feature space. Nonetheless, one or both of them could be discarded later on due to another pairwise evaluation. Indeed, when all the local correlations of a feature pair are equal or greater than the corresponding local thresholds (in panel B of Fig. 1, see the AND predicate on local similarity), just one of the two features is preserved. Specifically, the kept one is the feature having the highest local mean deviation, which is a differential measurement between two classes. This measure is the difference between the mean values of a feature in the two classes where such mean values are further away from each other, i.e. the deviation between the maximum and minimum class mean values for that feature.

Notably, the data-driven global and class-specific thresholds are the medians of the distributions of the corresponding pairwise feature correlations on the training set. Specifically, the values of every feature, respectively, in all training samples or in those samples belonging to a given class, are used to build an ordered vector, whose correlations to any other feature vector contribute to obtaining the distributions of the mentioned pairwise correlations. The choice of using median-based thresholds makes the ReRa approach more conservative than other Relevance-Redundancy strategies although the global similarity assessment is not very stringent: consequently, most of the decisions are taken based on supervised local similarity assessments.

2.2. *Benchmark of feature selection methods*

As feature selection benchmarks to compare with ReRa, we considered a set of canonical filters (also adopted within the relevance-based ReRa first step), LASSO regularization, and MRmr (as an alternative Relevance-Redundancy approach). On each feature space found with a different feature selection method, we used Logistic Regression, linear and polynomial Support Vector Machines, and Random Forest models [18–20] to perform the same desired classification and evaluate the suitability of the feature space under exam for the task.

2.2.1. *Filter methods*

Two knowledge-based and two statistical-based filters were adopted as feature selection benchmarks.

PAM50 [21] and LIMMA50 [22] gene signatures, including 50 and 257 genes, respectively, are two examples of filtering based on prior knowledge about the specific classification task: both signatures are indeed adopted for gene-level BRCA stratification into clinically-relevant subtypes [9,10]. PAM50 signature is used in the well-recognized namesake PAM50 subtyping test [21], while LIMMA50 signature has been recently traced [22] for the same stratification task since it includes genes differentially expressed among BRCA subtypes.

Alternatively, two univariate statistical filters were used to extract the most relevant features: Fisher Score [23] and Mutual Information [24]. Fisher Score is a measure of class separation, calculated considering the mean and standard deviation values of a given feature for each of the considered classes, and computing the sum across pairwise classes of the class mean differences divided by their standard deviation sum. Mutual Information computes the mutual dependence between two variables by measuring the reduction in uncertainty about one variable given the knowledge of the other one. Thus, it can also be used to evaluate the dependence of each feature on the target variable. Based on these statistical measures, we preserved all the features whose scores were above a given percentile of the Fisher Score or Mutual Information distribution, respectively.

Notice that each of the four considered filters was also used as an alternative first step of our ReRa approach, meant to select only the features of higher relevance for the classification task.

2.2.2. *Alternative feature selection methods*

The LASSO regularization was considered and compared with our ReRa approach; this embedded feature selection method is widely adopted since it can select a reduced set of predictive features within the training data. Indeed, a linear model is fitted with a LASSO regularization term: this shrinks the absolute values of the coefficients and sets many of them to zero, discarding the corresponding features. The preserved features can instead be used as a pre-filtered space to train further models, as we did in this work.

Furthermore, we adopted an already existing maximum relevance minimum redundancy approach (i.e., MRmr). The MRmr method, first proposed by Ding et al. [6], is a greedy search algorithm that iteratively selects the most relevant and non-redundant features from a dataset based on a combination of statistical measures to estimate feature relationships with the target variable and with any other feature. The MRmr algorithm here evaluated uses the analysis of variance (ANOVA) statistical test to calculate the F-statistic as the ratio between the variance explained by the predictor and the unexplained one (error variance) [25]. ANOVA is a parametric hypothesis test that evaluates any difference of means from two or more groups, using the variances to determine whether the means come from the same distribution. It can also work when one variable is numerical, like a quantitative feature, and one is categorical, like a classification target. The results of this test can be used for feature selection: those features that are independent of the target variable can be removed from the space of interest. The MRmr algorithm begins by selecting the feature with the highest F-statistic with the target variable and then adds additional features, one at a time, choosing at each step the feature having the highest F-statistic with the target variable and the minimum linear correlation with the already selected features. This selection proceeds until the desired number of features, which has to be predefined, is reached.

3. Application example: Clinically relevant genomics stratification of breast cancer patients

To show the relevance of our ReRa feature selection compared with other state-of-the-art methods, we applied it in the complex translational context of breast cancer patient stratification into clinically relevant target classes [10,21], based on patient gene or transcript isoform expression as two example use cases. Particularly, transcript isoforms are different forms of a gene RNA transcript produced by the same gene but having different sequences or structures due to alternative splicing events and transcription outcomes. These two example use cases are of increasing complexity in terms of both classification task, and dimensionalities of the initial feature spaces (12,381 genes vs. 49,740 isoforms, respectively, after the same pre-preprocessing). Indeed, the target BRCA classes, known as intrinsic subtypes, are five classes (Basal, HER2-enriched, Luminal A, Luminal B and Normal-like) originally traced [9] and still clinically assigned based on gene expression profiles of BRCA patients [21]. This stratification not only recognizes heterogeneity in BRCA molecular traits but is very useful and widely adopted to discriminate good-expected prognoses from poor ones, helping physicians decide about post-surgery chemotherapy and more suitable treatment options. Nonetheless, to the best of our knowledge, an in-depth evaluation of the contribution of transcript isoforms to intrinsic subtyping has not been addressed yet, particularly not through broad, well-annotated isoform expression data used as feature space for reliable machine learning techniques. Isoform diversity is known to be tissue- and disease-specific, and such specificity in BRCA has already been related to hormonal-status heterogeneity [26,27], which is also reflected within intrinsic subtypes. Yet, although the stratification into intrinsic subtypes is crucial in the clinical handling of BRCA patients, its evaluation at the isoform level as a more complete molecular and prognostically-relevant classification of patients has barely been touched [28].

Table 1
Class distributions in the training (410 samples) and test (127 samples) sets extracted from G_TCGA and I_TCGA datasets. These distributions are equal in the two datasets since both datasets refer to the same 537 patients, but they include as features 12,381 genes or 49,740 transcript isoforms, respectively.

Set	Basal	HER2-enriched	Luminal A	Luminal B	Normal-like
Training set	66 (16.1%)	33 (8.1%)	206 (50.2%)	92 (22.4%)	13 (3.2%)
Test set	24 (18.9%)	11 (8.7%)	64 (50.4%)	22 (17.3%)	6 (4.7%)
All data	90 (16.8%)	44 (8.2%)	270 (50.3%)	114 (21.2%)	19 (3.5%)

3.1. BRCA patient expression profiles: Datasets and pre-processing

Breast cancer patient cohorts are commonly stratified through the PAM50 test, a centroid-based subtyping method that focuses on the expression levels of 50 genes, known as the PAM50 signature. Assignments predicted by the PAM50 test are reported in the clinical data of the publicly disclosed BRCA patient dataset belonging to The Cancer Genome Atlas (TCGA) project [29], which includes gene expression profiles obtained from RNA-sequencing experiments. Additionally, for the majority of this patient cohort, transcript isoform expression profiles were also produced with RNA-sequencing experiments and are available to investigate. Therefore, the analyses in our example use cases are focused on the 537 TCGA patients for which both gene and transcript isoform expression profiles are at disposal.

Two datasets were considered: the gene-level one (G_TCGA), initially including 537 samples and 20,440 genes/features, and the isoform-level one (I_TCGA), initially including 537 samples and 67,347K transcript isoforms/features. For both datasets, Counts Per Millions (CPM) expression values were derived from the whole set of RNA-sequencing raw read counts available; then, the following pre-processing steps were performed:

- The features (genes or transcript isoforms) with null raw count value in at least 80% of the dataset samples were discarded from the analysis
- A $\log_2$ transformation was applied on the CPM data
- Too lowly expressed features were further discarded: only individual features whose median CPM value across all the samples was greater than the first quartile of the distribution of the median CPM values of all the features were kept.

After this pre-processing, the feature spaces of our datasets included 12,381 genes for G_TCGA and 49,740 isoforms for I_TCGA, respectively. Furthermore, each dataset was split into training (410 samples, 76.35%) and test (127 samples, 23.65%) sets, both reflecting the uneven class distributions of BRCA subtypes (see Table 1); these sets had very similar, although not identical, class proportions, due to the critical handling of the smallest classes (i.e., Normal-like and HER2-enriched). This allowed assessing the classification performance in a scenario with class occurrences likely real. Eventually, multiple feature spaces were obtained with different feature selection approaches and used to tune and train alternative classifiers on the training samples and evaluate classification results on the testing samples.

3.2. Supervised models for gene- and isoform-level subtyping

Different well-established machine learning classification techniques, usually advisable for classification tasks involving complex datasets of limited size, were compared in performing the same BRCA subtyping task when trained on ReRa-based predictive features or feature spaces from benchmark selection approaches: Logistic Regression, Support Vector Machines (SVM) with linear and polynomial kernels, and Random Forest. For every feature space under consideration, each model was trained supervisely using known subtyping class labels as targets and 10-fold cross-validation to limit the overfitting risk, while an extensive grid search allowed tuning the needed hyperparameters. Notice that the balanced accuracy was the chosen metric to be optimized, so as not to bias evaluations towards results mostly based on the contributions of the more numerous classes.

Each model was tested over the left-aside samples of the test set: specifically, given the multi-class context, we evaluated both macro- and micro-averaged performance measures. In macro-averaging, metrics are calculated independently for each class and then averaged across all classes; this is useful to make them contribute equally to the result even when classes are unbalanced (as in our case). Conversely, in micro-averaging, metrics are calculated globally by counting the number of true positives, false positives, and false negatives across all the classes; this is useful in unbalanced scenarios to give the same weight to each assignment (i.e., tested sample) regardless of its class. Yet, micro-averaged precision and recall in single-label classification contexts are both equal to the overall accuracy; consequently, the F1-score, their harmonic mean, is also coincident with the overall accuracy. Therefore, in the comparative result tables, we reported the overall accuracy as a micro-averaged measure, the balanced accuracy (i.e., the macro-averaged recall), the macro-averaged precision, and the macro-averaged F1-score (i.e., the harmonic mean of the macro-averaged precision and recall). This allowed us to detect any classification improvement due to our ReRa approach and to demonstrate its wide usability with alternative relevance-based strategies and classifiers.

3.3. Feature importance analysis and translational value assessment

From the two comprehensive evaluations of multiple different classifiers and feature spaces, we found the two most promising methods to perform BRCA subtyping at the gene- and transcript isoform-level, respectively. The two correspondent feature spaces were investigated using the Shapley Additive Explanations (SHAP) technique [30] for model interpretability.

Differently from other model-agnostic approaches like Permutation of Feature Importance [17] or Local Interpretable Model-agnostic Explanations (LIME) [31], the SHAP method computes the Shapley value of each feature, a measure from cooperative game theory that indicates how much a feature collaborates to obtain a given class prediction score. Specifically, this iterative method for feature importance analysis evaluates each feature based on its contribution to the prediction of all the testing instances of a given class, considering for each of them all possible combinations of features, including and excluding the specific feature under exam. In this way, it is possible to estimate quantitatively the importance of each feature to identify a given class and the consequences of its absence, also accounting for feature interactions. In addition, it is possible to estimate also the overall impact of a given feature on the whole classification task as the sum across all the classes of its mean SHAP values, each computed by averaging the feature absolute SHAP scores of all the testing samples assigned by the trained model to a specific class.

In the end, the features with the most crucial role at the gene- and transcript isoform-level were extracted and compared to preliminarily investigate the translational value of the found results, and to better understand any isoform contribution in distinguishing BRCA intrinsic subtypes and their clinically-relevant differences.

4. Results

For each of the two example use cases, the results obtained by the aforementioned classifiers on every assessed feature space are reported and discussed in the following subsections, focused on BRCA subtyping application at the gene- or isoform-level, respectively. Such results clearly demonstrate the efficacy of the proposed ReRa approach compared to the other feature selection methods described in Section 2.2.

4.1. Example use case on gene-level feature space

ReRa approach and benchmark feature selection methods were used to extract different feature spaces of interest from the 410 samples of the G_TCGA training set, each containing the expression values of the 12,381 pre-processed genes (i.e., features). Specifically, among these genes, 49 (out of 50) of the PAM50 signature and 185 (out of 257) of the LIMMA50 signature were available and sufficiently expressed to be used as two different feature spaces. By applying Fisher Score and Mutual Information measures and preserving the 97th, 98th or 99th percentile of their provided feature rankings, six additional feature spaces were extracted. Each of these 8 filtered feature spaces is also an intermediate result of the first step of the ReRa approach when using the corresponding knowledge- or filter-based strategy to maximize the extracted feature relevance; from them, 8 additional feature spaces were generated after completing the ReRa selection procedure with the feature redundancy minimization step.

Using MRmr, we extracted other 3 feature spaces, including 150, 250, or 350 genes, respectively, whose sizes are on purpose comparable to those selected with filter-based approaches (see the main diagonal of the table in Supplementary File F1 — Gene Tab); bigger sizes would indeed require unnecessarily higher computational costs and would also be more prone to keep some redundancy and noise. Finally, LASSO preserved a wider feature space of 917 genes, being a task-related selection method not strictly focused on feature redundancy minimization.

The amounts of genes included in each feature space and in the overlaps between pairs of feature spaces are all indicated in Supplementary File F1 (Gene Tab), together with their percentage with respect to the initial set of 12,381 genes; most of the independent feature spaces are barely overlapping, except those selected by MRmr and ReRa with Mutual Information, which suggests a certain consistency between linear and non-linear dependencies respectively evaluated by the two methods.

Performances of machine learning models trained on each of the feature spaces selected with filters or ReRa can be compared with those obtained with LASSO or MRmr, as to examine the suitability of each selection strategy to deal with the task and the features under consideration. Furthermore, feature dimensionality and classification performance on a final ReRa feature space can be compared with those of the corresponding intermediate filtered feature space to assess the value of the ReRa approach, particularly of its supervised step of feature redundancy minimization. To this aim, Table 2 shows as many rows referring to final feature spaces obtained with ReRa as for the 8 filtered feature spaces based only on feature relevance (each pair is reported in two consecutive rows of the table).

The performances collected for a given type of classifier were therefore evaluated either overall, i.e., against any other model, including those trained on LASSO or MRmr feature spaces, or pairwisely, i.e., comparing any pair of models trained on the intermediate or corresponding final feature spaces. The first evaluation is globally relevant: comparing different classifier types allows tracing the best solution according to the combination of use case, feature selection strategy and classifier. All pairwise comparisons, instead, highlight any effect due to the ReRa redundancy handling compared with the simple use of filtering. This allows us to evaluate our ReRa approach and its generalization capability, regardless of the chosen classification model or feature relevance metric maximized in the ReRa first step.

4.1.1. Computational performance and comparison with the state-of-the-art feature selection

In Table 2, we reported the testing performance of every SVM with a polynomial kernel properly tuned and trained on each considered feature space. In addition, when using statistical filters, Table 2 indicates the percentile chosen to extract the features, and, for each ReRa feature space, it also reports the percentage of feature reduction with respect to the corresponding feature space filtered only based on feature relevance.

Out of all the models and kernels explored for the considered classification task, polynomial SVM models appeared the most suitable overall to perform BRCA subtyping when we focused on our benchmark feature spaces from filters, LASSO, or MRmr. Also, polynomial SVMs proved the strengths of our ReRa approach, highlighting the contribution of its innovative similarity-based analysis in handling feature redundancy. Anyway, the main comments and evidence arising from the analysis of Table 2 can be straightforwardly extended to the other classification models: this emerges clearly from Supplementary File F2 (Gene Tab), which lists feature dimensions, tuned values of the main hyperparameters and testing performances for all the classifiers trained on each of the compared gene-level feature spaces.

First of all, we can notice that stronger filter-based feature selections mostly bring better results than the commonly-used LASSO approach; the only exception is given by the Fisher Score-based filtering, which leads to under-performing results for whichever derived feature space and chosen classification model (Supplementary File F2), suggesting that this statistical measure is not adequate to retrieve significantly predictive features for this specific task. Nevertheless, even when considering such disadvantageous Fisher Score-based feature spaces and the classification models trained on them, in more than 80% of the cases, the performances are enhanced when using our ReRa approach compared to the simple filtering (Supplementary File F2). Improvements indicate that our ReRa feature spaces allow obtaining similar or higher (in the vast majority of the cases) performances compared to statistical filters, using a further reduced number of more predictive features. Such improvements concern over 90% of the collected pairwise comparisons in our gene-level investigation, particularly all the cases using knowledge-based and Mutual Information filtering. Since each classifier is tuned and trained independently from all the others, we can observe slight changes in the trade-off between macro- and micro-averaged metrics, even for pairs of models trained on corresponding intermediate and final feature spaces. Nevertheless, all pairwise comparisons based on Mutual Information or knowledge-based signatures demonstrate that our ReRa method can enhance the gene-level subtyping classification compared to canonical filters, besides greatly outperforming the models trained on LASSO feature spaces.

Results obtained with ReRa using knowledge-based filters are mostly comparable with those using the Mutual Information measure. Yet, these latter ones provided slightly higher performance, particularly when considering the 97th or 98th percentile cut. Conversely, for greater feature amounts (data not shown) both filter- and ReRa-based results tend to stability or slight decrease, suggesting to focus just on the considered percentiles: these allow working on restrained sets of relevant features able to guarantee better performance and computational efficiency when applying the complete ReRa approach.

Regarding MRmr feature selection, good performances were reached all over the assessed feature space cardinalities and considered models, except for the simple Logistic Regression, which appeared slightly weaker. Both overall and balanced accuracies were instead particularly valuable when using polynomial SVMs and Random Forests on such feature spaces. Using this latter ensemble method, our results based on ReRa feature spaces were comparable to those on MRmr feature spaces. In contrast, the other ReRa-based classifiers greatly outperformed MRmr-based ones almost always, also when focusing on very different feature sizes. Such results are even more interesting considering the number of overlapping genes among the feature spaces originating using the Mutual Information measure and those obtained with MRmr (see Supplementary File F1 — Gene Tab). The ReRa approach based on Mutual Information selected in a more efficient way many predictive features confirmed by MRmr; in addition, ReRa was able to trace other key features leading towards better performance

Table 2
Feature space dimensionalities and performance evaluations on the test set for each SVM with polynomial kernel tuned and trained on gene-level feature spaces.

Feature selection	Percentile for selection	Kept features	Feature reduction	Balanced accuracy	Overall accuracy	Precision macro-avg	F1-score macro-avg
LASSO selection	–	917	–	0.75	0.85	0.80	0.76
MRmr selection	–	150	–	0.77	0.88	0.89	0.81
MRmr selection	–	250	–	0.82	0.88	0.82	0.82
MRmr selection	–	350	–	0.78	0.87	0.81	0.79
PAM50 filter	–	49	–	0.79	0.86	0.80	0.79
ReRa (with PAM50)	–	41	16.3%	0.81	0.86	0.79	0.80
LIMMA50 filter	–	185	–	0.80	0.89	0.85	0.82
ReRa (with LIMMA50)	–	154	16.8%	0.83	0.90	0.84	0.83
Fisher Score filter	99%	124	–	0.65	0.78	0.68	0.65
ReRa (with Fisher Score)	99%	84	32.3%	0.63	0.79	0.76	0.63
Fisher Score filter	98%	248	–	0.70	0.82	0.74	0.71
ReRa (with Fisher Score)	98%	169	31.9%	0.71	0.81	0.71	0.71
Fisher Score filter	97%	372	–	0.65	0.79	0.74	0.67
ReRa (with Fisher Score)	97%	248	33.3%	0.65	0.80	0.73	0.67
Mutual Information filter	99%	124	–	0.79	0.87	0.87	0.81
ReRa (with Mutual Info.)	99%	116	6.5%	0.82	0.91	0.92	0.85
Mutual Information filter	98%	248	–	0.80	0.88	0.89	0.82
ReRa (with Mutual Info.)	98%	212	14.5%	0.82	0.91	0.92	0.85
Mutual Information filter	97%	372	–	0.83	0.90	0.86	0.84
ReRa (with Mutual Info.)	**97%**	**320**	**14.0%**	**0.85**	**0.89**	**0.88**	**0.86**

compared to MRmr: this corroborates the value of our proposed feature selection strategy.

Finally, from a wider view of this application use case, we can observe that the predictions of any classifier cannot perfectly reconstruct the target assignments; this is expected for such a complex subtyping task since target assignments do not depend on indisputable ground truth, but on a gold standard method (PAM50) that comes with its inherent limitations [22]. Nonetheless, for Logistic Regression and Random Forest models (Supplementary File F2 — Gene Tab), ReRa-based classifiers appeared enhanced in 100% of the cases, although these models show lower subtyping capabilities than SVMs. Conversely, polynomial SVMs show the most significant performances across all the feature spaces, including benchmark feature spaces; this could have left smaller room for performance improvement brought by our ReRa redundancy evaluation step. In contrast, polynomial SVMs trained on ReRa feature spaces using Mutual Information as a relevance measure not only outperform their counterparts trained on just filtered feature spaces but are also among the top most promising classifiers, reaching the highest overall and balanced accuracies.

Particularly, the best method for gene-level BRCA subtyping was the polynomial SVM trained on the ReRa feature space of 320 features selected within the 97th percentile of the most relevant features based on Mutual Information. After discarding 52 redundant features while preserving those differentiated within classes, this polynomial SVM reached the maximum balanced accuracy of 0.85, with both increased precision and F1-score, and almost unvaried overall accuracy. The remarkable result found by this model on such ReRa feature space was investigated with an accurate feature importance analysis, able to estimate the contribution of each feature to the predictive task.

4.1.2. Comparison against random feature spaces and translational evidences

To further prove the value of our ReRa feature selection, the ReRa feature space obtained from the 97th percentile of the Mutual Information-based feature ranking, on which we found the most performing polynomial SVM, was compared to ten random feature selections (with as many features), used to tune and train polynomial SVMs. Despite having exactly the same amount of 320 features, these random feature spaces scarcely overlapped with our ReRa-based feature space (2.59 ± 0.81%). The high information redundancy within the initial gene set allowed collecting barely marginal performances on such random feature spaces, whose results were much lower than those gained using our ReRa feature space, with balanced accuracy values approximately 20% smaller.

The features of such best ReRa feature space were also explored with the computation of their Shapley values (reported in Supplementary File F3), as to identify the genes with the most crucial roles in BRCA subtyping, either overall or for each specific BRCA class. All top 10 most relevant genes overall were also in the top 10 genes of at least one specific class: 3/10 were among the most relevant for the Basal cases, 4/10 for HER2-enriched, 7/10 for Luminal A, 8/10 for Luminal B, and 2/10 for Normal-like. Although many genes resulted crucial for more classes, the contribution of each gene for recognizing one class may be stronger than for other classes: accordingly, gene orderings and scores in the six assessed rankings were completely different from one another.

In Supplementary File F3, we also reported the functional enrichment analysis results obtained by examining the overall top ten genes against the pathways in KEGG, Reactome and Wikipathways databases. The significantly enriched annotations were evaluated based on the involved genes and their presence in class-specific rankings to infer relationships among different classes and molecular phenomena. Most of the pathways resulted associated with the majority of the classes, as expected for ’Breast Cancer pathway’ or annotations referring to Estrogen or ERBB4 signaling; yet, some of them highlighted particular traits for only a few subtypes. For instance, both RUNX1 regulation and the WNT signaling pathway appeared to be associated with the Basal and HER2-enriched subtypes of BRCA. RUNX1 transcription factor and WNT signaling pathway are critical in regulating cell proliferation and differentiation: in Basal BRCA, RUNX1 decreased expression and loss of function as well as increased WNT pathway activation have been widely associated with more aggressive disease and poorer outcomes [32–34]. In HER2-enriched BRCA, although the specific role of RUNX1 is less clear, the WNT pathway is activated, suggesting that it may play a key role in this subtype development and progression [35]. Some other annotations still miss clear supporting evidence, like the Aflatoxin B1 (AFB1) metabolism that may be associated with Luminal A and Luminal B BRCA subtypes. Exposure to AFB1 has been shown to cause epigenetic changes leading to carcinogenesis, especially in hepatocellular carcinoma [36], and the role of epigenetic modifications in endocrine treatment resistance of Luminal breast cancer is under study [37,38]. Thus, these highlighted associations may be worthy

Table 3
Feature space dimensionalities and performance evaluations on the test set for each SVM with polynomial kernel tuned and trained on isoform-level feature spaces.

Feature selection	Percentile for selection	Kept features	Feature reduction	Balanced accuracy	Overall accuracy	Precision macro-avg	F1-score macro-avg
LASSO selection	–	1767	–	0.75	0.87	0.83	0.79
MRmr selection[a]	–	500	–	0.81	0.89	0.85	0.82
MRmr selection[a]	–	750	–	0.79	0.87	0.83	0.81
MRmr selection[a]	–	1000	–	0.80	0.88	0.84	0.82
PAM50 filter	–	131	–	0.80	0.87	0.79	0.79
ReRa (with PAM50)	–	129	1.5%	0.83	0.86	0.79	0.81
LIMMA50 filter	–	557	–	0.75	0.87	0.89	0.78
ReRa (with LIMMA50)	–	533	4.3%	0.78	0.85	0.81	0.79
Fisher Score filter	99%	498	–	0.61	0.78	0.60	0.60
ReRa (with Fisher Score)	99%	393	21.1%	0.57	0.76	0.55	0.57
Fisher Score filter	98%	995	–	0.63	0.80	0.61	0.62
ReRa (with Fisher Score)	98%	792	20.4%	0.66	0.82	0.64	0.65
Fisher Score filter	97%	1493	–	0.65	0.82	0.64	0.64
ReRa (with Fisher Score)	97%	1208	19.1%	0.64	0.81	0.62	0.63
Mutual Information filter	99%	498	–	0.80	0.88	0.89	0.83
ReRa (with Mutual Info.)	99%	440	11.6%	0.80	0.88	0.89	0.83
Mutual Information filter	98%	995	–	0.81	0.90	0.90	0.84
ReRa (with Mutual Info.)	**98%**	**887**	**10.9%**	**0.85**	**0.88**	**0.88**	**0.86**
Mutual Information filter	97%	1493	–	0.81	0.90	0.87	0.81
ReRa (with Mutual Info.)	97%	1264	15.3%	0.82	0.87	0.82	0.81

[a] Due to its computational demand, MRmr is executed on the top 12k most expressed isoforms.

of further investigations to be confirmed, and to inspect any possible translational value for clinical handling or therapeutic design.

4.2. Example use case on isoform-level feature space

From the 410 samples of the I_TCGA training set, with their initial 49,740 transcript isoforms (i.e., features), we extracted the feature spaces of interest to train and test supervised models, analogously to what was done for the gene-level investigation. We selected all available transcript isoforms originating from genes belonging to the PAM50 or LIMMA50 gene signatures: the so-derived PAM50 feature space included 131 isoforms, while the LIMMA50 one contained 557 isoforms. Statistical filtering (based on Fisher Score or Mutual Information) was directly applied on the initial 49,740 transcript isoforms to extract six additional feature spaces, considering again the 97th, 98th and 99th percentile of the feature rankings. On each of these 8 feature spaces filtered based on relevance strategies, the ReRa step of redundancy minimization was applied to obtain corresponding final ReRa feature spaces. LASSO regularization automatically extracted 1,767 isoforms from the initial feature space, while MRmr was used to select feature spaces including 500, 750, or 1,000 isoforms, equivalent to the dimensionalities of the statistical and ReRa-based selections. The amounts of isoforms/features selected by each method are reported in Supplementary File F1 (Isoform Tab), which also includes the feature overlaps between each pair of feature spaces.

4.2.1. Computational performance and comparison with the state-of-the-art feature selection

For every classifier and each of the 8 ReRa isoform feature spaces extracted, we evaluated performance comparisons against correspondent classification models trained on LASSO, MRmr, or filtered-only feature spaces. Thus, we assessed the fitness of every filtering strategy for the specific task and features under exam, as well as the value of the ReRa step of supervised similarity-based assessments to minimize selected feature redundancy. Table 3 shows the main results of these comparisons when focusing on the polynomial SVM models, which still dominate the other classifiers; the complete set of outcomes is provided in the Isoform Tab of Supplementary File F2.

Even at the isoform-level, any type of knowledge-based and Mutual Information-based feature spaces provides each machine learning model with fewer but more crucial features for class distinction than those selected by LASSO, leading to better subtyping performances in testing. Still, our ReRa feature spaces further enhanced the subsequent classification compared to corresponding statistical filters for more than 85% of these cases, letting classification models obtain similar or mainly higher performances but using a reduced feature amount.

Notice that selected feature sizes are much higher than those preserved for gene-level subtyping, although the percentiles of interest are the same: indeed, the initial isoform-level feature space is more than 4 times bigger than the gene-level one, and both inner variability and redundancy risk are greater at the isoform-level. Accordingly, while feature spaces selected using Mutual Information only for relevance-based filtering include from five hundred to one thousand and five hundred isoforms, the corresponding ReRa-based ones are reduced by 10%–15%; this is due to the minimization step that optimizes the performances while automatically controlling global and local redundancy.

Fisher Score-based feature spaces appeared inadequate to perform the subtyping task also using isoforms. This affected the corresponding ReRa-based spaces (approximately 20% smaller in size than their Fisher-based filtered-only counterparts), more reduced than those based on Mutual Information, but providing just slight performance enhancements in barely one-half of the cases across all classification models.

However, as in gene-level analysis, Random Forest models are always enhanced by the use of ReRa feature spaces: for both investigation levels, this kind of model reaches a high performance of overall accuracy, although it is slightly penalized in terms of balanced accuracy; notably, both these metrics are relevant to be assessed in the case of uneven class distribution.

Overall, all pairwise comparisons based on Mutual Information or knowledge-based feature spaces demonstrate that our ReRa method, once chosen a relevance strategy suitable for the task, is able to enhance the classification also at the isoform-level, compared to filters, LASSO, or MRmr feature selection. Furthermore, ReRa-based models using Mutual Information and knowledge-based signatures mostly outperformed the models trained on MRmr feature spaces, even when these latter include comparable feature amounts. Thus, classification results confirm the capability of our ReRa approach to provide key features for class distinction, especially when using the Mutual Information metric, regardless of the filtering percentile and chosen model, with clear subtyping improvements (see Supplementary File F2).

Despite subtype predictions cannot perfectly reconstruct the target assignments, the polynomial SVMs trained on ReRa feature spaces (based on Mutual Information or PAM50) were again the most performing models overall. Specifically, the model with the most remarkable result for isoform-level BRCA subtyping was the polynomial SVM trained on the ReRa 887-feature space, selected within the 98th percentile of the Mutual Information ranking. This ReRa feature space discarded 108 redundant features compared to its filtered-only counterpart and allowed the polynomial SVM to reach a balanced accuracy of 0.85, the same maximum value reached with gene-level subtyping. Therefore, such ReRa isoform feature space was accurately investigated to estimate feature importance and assess the contribution of each isoform to the subtyping task, also comparing this analysis with that performed at the gene level.

4.2.2. Comparison against random feature spaces and translational evidences

The ReRa 887-feature space coming from the 98th percentile of the Mutual Information-based feature ranking, (on which the best isoform-level subtyping result has been obtained) was compared to ten random selections of the same amount of isoforms, each one used to tune and train other polynomial SVMs. These random feature spaces had a very low overlap with our ReRa-based feature space (1.65 ± 0.33%) and made SVMs reach only marginal performances, with balanced accuracy values again 20% smaller than that obtained from the ReRa feature space. This further proved our ReRa feature selection reliability and value.

The isoforms of such ReRa feature space were also explored to identify those with the most crucial roles in BRCA subtyping, either overall or for each specific class. Their computed Shapley values, together with class-specific and overall rankings, are reported in Supplementary File F4. All top 10 most relevant isoforms overall also belong to the top 10 isoforms of at least one specific BRCA class: 1/10 in Basal class top ranking, 6/10 in HER2-enriched, 7/10 in Luminal A, 6/10 in Luminal B, and 4/10 in Normal-like. Similarly to gene-level BRCA subtyping, the orderings of the key features differed across the classes, even when these shared some of their relevant isoforms. In Supplementary File F4, we also reported the genes of origin for the most relevant top ten isoforms overall and of each class. For overall and Luminal A feature rankings, two isoforms originating from the same gene GFRA1 resulted within the top 10 positions; this also occurred for the HER2-enriched class with the gene ESR1 and for the Normal-like class with the gene GABRP, whilst both these genes were not among those encoding the most relevant isoforms overall.

Additionally, Supplementary File F4 includes the functional enrichment analysis results obtained by testing the genes of origin of the top ten overall isoforms against KEGG, Reactome and Wikipathways databases. As for gene-level subtyping, for each significantly enriched annotation referring to a specific subgroup of BRCA subtypes, we examined the involved isoform-related genes associated with that annotation. Interestingly, we found some significant annotations that highlight additional aspects that did not emerge from gene-level investigation. These comprise the GDNF/RET signaling pathway, which has been implicated in the development and progression of various types of cancer, including BRCA [39], and that may be associated with more aggressive disease in Luminal B cases, according to some recent evidence [40,41]. Similarly, additional annotations refer to the interactions of the NCAM1 cell adhesion molecule, which have been shown to regard various signaling pathways that are dysregulated in BRCA, especially in case of cancer migration to lymph nodes and worse expected clinical outcome [42]. Such annotations, which emerged only with the isoform-level investigation, reflect molecular phenomena that may be useful to improve the tough distinction of Luminal B from Luminal A (having better expected prognosis) BRCA patients, showing the value of isoform-level analysis. Comparison between the most relevant features and consistency of functional annotations arising from isoform- and gene-level BRCA subtyping are discussed in the following subsection.

4.3. Comparative evaluation of the two example use cases

From our gene-level and isoform-level investigations, using the ReRa approach, we retrieved two meaningful feature spaces that guarantee convincing BRCA subtyping performances, especially for polynomial SVM classifiers. These two feature spaces were compared considering for each transcript isoform the corresponding gene of origin: 244 of the 753 genes corresponding to the 887 isoforms of the best isoform-level ReRa feature space were in common with the 320 genes of the best feature space selected by ReRa during the gene-level investigation. Accordingly, more than 75% of the genes involved in BRCA subtyping at gene-level were confirmed and explained more in detail considering the isoform specificity, which is usually left out in such kinds of analyses. Furthermore, for approximately more than 5 hundred other isoforms (and their corresponding genes), we were able to estimate a quantitative contribution in BRCA subtyping and their specific relevance for the distinction of each subtype (Supplementary File F4); this further demonstrates the foreseen relevance of transcript isoform-level analysis [43–45], which our ReRa approach makes more computationally affordable.

Eventually, for what concerns functional enrichments, almost one-third of the significant annotations found based on the top ten features for the best gene-level or isoform-level BRCA subtyping models were detected at both subtyping levels. These include functional terms clearly associated with BRCA, like ’Estrogen signaling pathway’, ’Estrogen-dependent gene expression’ and ’Mammary Gland Development pathway’. However, several other enriched functional terms emerged as associated with only the top ten features from the gene-level or isoform-level investigation (see Supplementary File F3 vs F4); these provide different and interesting preliminary perspectives on the clinically relevant subtyping task under exam and could steer further studies, also useful to assess any possible translational value of the so-found insights.

5. Discussion

Through the illustrated application use cases and the results obtained from their evaluation, we proved the efficacy of our ReRa feature selection approach in two alternative translational scenarios for clinically-relevant breast cancer subtyping. Both of them are characterized by feature high dimensionality and redundancy, and by a limited number of available samples, with a strongly unbalanced class distribution. The ReRa approach allowed us to select feature spaces where different classifiers succeed in well distinguishing target classes despite their uneven proportions, as demonstrated by several performance metrics including balanced accuracy and other macro- and micro-averaged measures that inspect and aggregate the outcomes on each class.

In the vast majority of the assessed cases, the collected classification performances when using ReRa feature spaces are better than those reached using feature spaces from simple feature filtering, LASSO regularization, or the MRmr method. The comparison with this latter one is even more interesting considering that both ReRa and MRmr are alternative Relevance-Redundancy strategies, and that there is a good overlap among the ReRa feature spaces originating using the Mutual Information measure and those obtained with MRmr. While MRmr, which simultaneously evaluates Relevance and Redundancy for each feature, struggles to scale up on huge initial spaces, the ReRa approach has its relevance-based filtering step that ensures scalability and discards useless features, especially when working on big unbalanced datasets. To this aim, similarly to LASSO and differently from MRmr, ReRa evaluated both the here-considered initial spaces without requiring any tuning of the number of features to be kept. MRmr instead could not even be applied to the initial dimension of our isoform-level investigation (around 49k isoforms), requiring to decrease it to approximately one-fifth (based on the most expressed isoforms) to be run. Also, the time efficiency evaluations, reported in Supplementary File F5, clearly show that ReRa usually requires much less than half of the running time compared to MRmr to select a comparable subset of predictive features. Thus, the ReRa approach selects features in a more efficient way, regardless of the initial dimensionality.

In addition, ReRa relevance strategy is open to alternatives and generalizations. This is an added value since any classification task has supervised strategies of feature relevance estimation more suitable than others; therefore, it is preferable not to make prior assumptions, but rather to explore and compare several options (e.g., statistical vs. knowledge-based filters), as it is good practice also for classifier choice. However, the most relevant and innovative aspect of ReRa is given by its second step of supervised feature similarity assessments to minimize feature redundancy.

The similarity assessments are performed iteratively over the set of previously chosen relevant features and allow reevaluating and updating the selected feature space at each iteration; this is another difference from MRmr. Furthermore, ReRa evaluates both global and class-specific (local) feature redundancies, with each class being one of the targets of the predictive task under exam: this allows not ignoring any differential behavior that may be remarkable for better distinguishing a class, but hidden at the global level. Thus, besides the improvement brought in classification results, our novel ReRa strategy provides features offering interesting insights about class differentiation, worthy of further investigation, especially in the case of unbalanced class proportions.

This kind of strongly unbalanced scenario, with huge feature spaces and limited available samples, is very common in relevant biomedical applications that require dealing with omics data to assign patients to subgroups, based on their clinical/biological heterogeneity, to ensure personalized healthcare decisions. Performing patient stratification and highlighting the most predictive features can increase current knowledge and boost precision medicine. Particularly, the breast cancer subtyping task considered here to validate our ReRa approach is an insightful example of translational bioinformatics computational application. Investigating transcript isoforms as features for this clinically-relevant task and comparing isoform- and gene-level stratifications may have interesting translational implications: isoform differentiation, hidden at the gene level, could be a precious resource to better characterize, distinguish, and even treat different cancer subtypes. Transcript isoforms can indeed alter and discriminate functions and molecular products, and their further study could underline actionable drug targets. Here, we used our comparative survey not only to validate our ReRa approach but also to provide meaningful parallelism between gene- and transcript isoform-level subtyping based on machine learning models, particularly offering new knowledge about isoform role in subtype differentiation. Furthermore, the investigation performed here is well generalizable and can be easily applied also to other cancer types and diseases to dissect their heterogeneity at different molecular levels.

6. Conclusions

We proposed and comparatively evaluated our innovative feature selection methodology, ReRa, based on supervised Relevance-Redundancy assessments. We clearly proved ReRa efficacy in two wide comparative translational scenarios aiming to provide clinically-relevant BRCA patient stratification at gene- and transcript isoform-level. The two considered use cases represent an insightful example of translational application, taking advantage of ReRa capabilities to investigate and enhance a clinically-relevant patient stratification task, which could be easily applied also to other cancer types and diseases.

Overall, the innovative ReRa approach demonstrated to be efficient and scalable, comparable to filter-based and LASSO methods; similarly to this latter one, it uses an iterative optimization process for feature selection. While offering these strengths over the MRmr approach, ReRa is still a Relevance Redundancy strategy; therefore, ReRa can reduce the initial feature space to a more compact set of relevant and non-redundant features, after extensive examination of feature relationships both globally and locally to the target classes. In the vast majority of the assessed scenarios, when using ReRa-selected feature spaces the performances were significantly increased compared to simple filtering, LASSO, or MRmr feature selection. Without the need to tune the number of preserved features, ReRa reduces the number of features while ultimately maintaining the most relevant and class-differentiated ones. To this aim, it iteratively re-evaluates the previously chosen relevant features getting rid of global and, mostly, local redundancies found at the level of each class of interest. This peculiarity allows highlighting class-specific behaviors that can play a hidden crucial role in a classification task, particularly when unbalanced class distributions need to be tackled.

Thus, ReRa, with its two-step structure, is able to ensure efficiency and scalability over huge initial dimensionalities, while selecting a compact set of relevant and non-redundant predictive features. So-selected features offer better insights about class differentiation in highly unbalanced classification scenarios and lead to better performance of machine learning models, even in highly unbalanced classification scenarios. Lastly, ReRa wide applicability and generalization power corroborate its proven value as a supervised Relevance-Redundancy feature selection strategy.

CRediT authorship contribution statement

Silvia Cascianelli: Conceptualization, Methodology, Software, Writing – reviewing. **Arianna Galzerano:** Methodology, Software development and maintenance. **Marco Masseroli:** Conceptualization, Supervision, Writing – reviewing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank our former student Sotoodeh Naeemi Sanatdoost, who applied and tested the preliminary version of our ReRa approach.

Appendix A. Supplementary data

Supplementary material related to this article can be found online and includes the following files:

- Supplementary File F1 (Suppl_File_1.xlsx): Tables containing the amounts of features at gene-level (Gene Tab) or isoform-level (Isoform Tab) of each considered feature space (on the main diagonal) or in common (elsewhere) between a pair of feature spaces;
- Supplementary File F2 (Suppl_File_2.xlsx): Performances of all the trained classifiers on each of the assessed feature spaces at the gene-level (Gene Tab) or isoform-level (Isoform Tab);
- Supplementary File F3 (Suppl_File_3.xlsx): Feature importance and functional enrichment analyses based on the best ReRa feature space to perform gene-level subtyping of BRCA patients;

- Supplementary File F4 (Suppl_File_4.xlsx): Feature importance and functional enrichment analyses based on the best ReRa feature space to perform isoform-level subtyping of BRCA patients.
- Supplementary File F5 (Suppl_File_5.xlsx): Running time and number of input and selected features in several experiments using ReRa approach compared with the MRmr method.

The developed Python code with our ReRa implementation and its benchmarking is publicly available at https://github.com/DEIB-GECO/BRCA_ISOFORMS.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104457.

References

[1] J. Tang, Feature selection for classification: a review, in: Data Classification: Algorithms and Applications, Vol. 6, 2014, pp. 37–64.

[2] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28.

[3] P. Dhal, C. Azad, A comprehensive survey on feature selection in the various fields of machine learning, Appl. Intell. (2022) 1–39.

[4] R. Porkodi, Comparison of filter based feature selection algorithms: an overview, Int. J. Innov. Res. Technol. Sci. 2 (2) (2014) 108–113.

[5] A. Bommert, T. Welchowski, M. Schmid, J. Rahnenführer, Benchmark of filter methods for feature selection in high-dimensional gene expression survival data, Brief. Bioinform. 23 (1) (2022) bbab354.

[6] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, J. Bioinform. Comput. Biol. 3 (2) (2005) 185–205.

[7] S. Mandal, A. Mukhopadhyay, An improved minimum redundancy maximum relevance approach for feature selection in gene expression data, Proc. Technol. 10 (2013) 20–27.

[8] Z. Zhao, R. Anand, M. Wang, Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform, in: 2019 IEEE International Conference on Data Science and Advanced Analytics, DSAA, IEEE, 2019, pp. 442–452.

[9] 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 406 (6797) (2000) 747–752.

[10] X. Dai, T. Li, Z. Bai, Y. Yang, X. Liu, J.-W. Zhan, B. Shi, Breast cancer intrinsic subtype classification, clinical use and future trends, Am. J. Cancer Res. 5 (10) (2015) 2929–2943.

[11] R.J. Urbanowicz, R.S. Olson, P. Schmitt, M. Meeker, J.H. Moore, Moore, Benchmarking relief-based feature selection methods for bioinformatics data mining, J. Biomed. Inform. 85 (2018) 168–188.

[12] J. Dai, J. Chen, Feature selection via normative fuzzy information weight with application into tumor classification, Appl. Soft Comput. 92 (2020) 106299.

[13] P. Yildirim, Filter based feature selection methods for prediction of risks in hepatitis disease, Int. J. Mach. Learn. Comput. 5 (4) (2015) 258.

[14] N. El Aboudi, L. Benhlima, Review on wrapper feature selection approaches, in: 2016 International Conference on Engineering & MIS, ICEMIS, IEEE, 2016, pp. 1–5.

[15] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 267–288.

[16] R.J. Urbanowicz, M. Meeker, W. La Cava, R.S. Olson, J.H. Moore, Relief-based feature selection: Introduction and review, J. Biomed. Inform. 85 (2018) 189–203.

[17] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (10) (2010) 1340–1347.

[18] I. Steinwart, A. Christmann, Support Vector Machines, Springer Science & Business Media, 2008.

[19] D.G. Kleinbaum, K. Dietz, M. Gail, M. Klein, M. Klein, Logistic Regression, Springer, 2002.

[20] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[21] J.S. Parker, M. Mullins, M.C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, et al., Supervised risk predictor of breast cancer based on intrinsic subtypes, J. Clin. Oncol. 27 (8) (2009) 1160.

[22] S. Cascianelli, I. Molineris, C. Isella, M. Masseroli, E. Medico, Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer, Sci. Rep. 10 (1) (2020) 1–13.

[23] M.R. Osborne, Fisher's method of scoring, Int. Stat. Rev./Revue Internationale de Statistique (1992) 99–117.

[24] T.M. Cover, Elements of Information Theory, John Wiley & Sons, 1999.

[25] R.A. Fisher, Statistical Methods for Research Workers, Springer, 1992.

[26] T.P. Stricker, C.D. Brown, C. Bandlamudi, M. McNerney, R. Kittler, V. Montoya, A. Peterson, R. Grossman, K.P. White, Robust stratification of breast cancer subtypes using differential patterns of transcript isoform expression, PLoS Genet. 13 (3) (2017) e1006589.

[27] A. Read, R. Natrajan, Splicing dysregulation as a driver of breast cancer, Endocrine-related Cancer 25 (9) (2018) R467.

[28] L. Wang, Y. Wang, B. Su, P. Yu, J. He, L. Meng, Q. Xiao, J. Sun, K. Zhou, Y. Xue, et al., Transcriptome-wide analysis and modelling of prognostic alternative splicing signatures in invasive breast cancer: a prospective clinical study, Sci. Rep. 10 (1) (2020) 1–16.

[29] J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R.M. Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J.M. Stuart, The Cancer Genome Atlas pan-cancer analysis project, Nature Genet. 45 (10) (2013) 1113–1120.

[30] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[31] M.T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning, 2016, arXiv preprint arXiv:1606.05386.

[32] N. Dey, B.G. Barwick, C.S. Moreno, M. Ordanic-Kodani, Z. Chen, G. Oprea-Ilies, W. Tang, C. Catzavelos, K.F. Kerstann, G.W. Sledge, et al., Wnt signaling in triple negative breast cancer is associated with metastasis, BMC Cancer 13 (1) (2013) 1–15.

[33] N. Ferrari, Z.M. Mohammed, C. Nixon, S.M. Mason, E. Mallon, D.C. McMillan, J.S. Morris, E.R. Cameron, J. Edwards, K. Blyth, Expression of RUNX1 correlates with poor patient prognosis in triple negative breast cancer, PLoS One 9 (6) (2014) e100759.

[34] N.B. Fernández, S.M. Sosa, J.T. Roberts, M.S. Recouvreux, L. Rocha-Viegas, J.L. Christenson, N.S. Spoelstra, F.L. Couto, A.R. Raimondi, J.K. Richer, et al., RUNX1 is regulated by androgen receptor to promote cancer stem markers and chemotherapy resistance in triple negative breast cancer, Cells 12 (3) (2023) 444.

[35] N. Mukherjee, N. Bhattacharya, N. Alam, A. Roy, S. Roychoudhury, C.K. Panda, Subtype-specific alterations of the wnt signaling pathway in breast cancer: clinical and prognostic significance, Cancer Sci. 103 (2) (2012) 210–220.

[36] S.K. Mungamuri, V.A. Mavuduru, Role of epigenetic alterations in aflatoxin-induced hepatocellular carcinoma, Liver Cancer Int. 1 (2) (2020) 41–50.

[37] H.A. Abdel-Hafiz, K.B. Horwitz, Role of epigenetic modifications in luminal breast cancer, Epigenomics 7 (5) (2015) 847–862.

[38] F.-I. Dimitrakopoulos, A. Kottorou, A. Tzezou, Endocrine resistance and epigenetic reprogramming in estrogen receptor positive breast cancer, Cancer Lett. 517 (2021) 55–65.

[39] S. Esseghir, S.K. Todd, T. Hunt, R. Poulsom, I. Plaza-Menacho, J.S. Reis-Filho, C.M. Isacke, A role for glial cell–derived neurotrophic factor–induced expression by inflammatory cytokines and RET/GFR α1 receptor up-regulation in breast cancer, Cancer Res. 67 (24) (2007) 11732–11741.

[40] R. Mechera, S.D. Soysal, S. Piscuoglio, C.K. Ng, J. Zeindler, E. Mujagic, S. Däster, P. Glauser, H. Hoffmann, E. Kilic, et al., Expression of RET is associated with Oestrogen receptor expression but lacks prognostic significance in breast cancer, BMC Cancer 19 (1) (2019) 1–10.

[41] A.C. Pavanelli, F.R. Mangone, P. Yoganathan, S.A. Bessa, S. Nonogaki, C.A. de Toledo Osório, V.P. de Andrade, I.C. Soares, E.S. de Mello, L.M. Mulligan, et al., Comprehensive immunohistochemical analysis of RET, BCAR1, and BCAR3 expression in patients with Luminal A and B breast cancer subtypes, Breast Cancer Res. Treat. (2022) 1–10.

[42] S. Dorman, C. Viner, P. Rogan, Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer, Sci. Rep. 4 (7063) (2014) 1–9.

[43] Z. Zhang, S. Pal, Y. Bi, J. Tchou, R.V. Davuluri, Isoform level expression profiles provide better cancer signatures than gene level expression profiles, Genome Med. 5 (2013) 1–13.

[44] E. Sebestyén, M. Zawisza, E. Eyras, Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer, Nucleic Acids Res. 43 (3) (2015) 1345–1356.

[45] N.T. Johnson, A. Dhroso, K.J. Hughes, D. Korkin, Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers? RNA 24 (9) (2018) 1119–1132.