

Learning From High-Dimensional Biomedical Datasets: The Issue of Class Imbalance

BARBARA PES^{ID}, (Member, IEEE)

Dipartimento di Matematica e Informatica, Università degli Studi di Cagliari, 09124 Cagliari, Italy

e-mail: pes@unica.it

ABSTRACT As witnessed by a vast corpus of literature, dimensionality reduction is a fundamental step for biomedical data analysis. Indeed, in this domain, there is often the need for coping with a huge number of data attributes (or features). By removing irrelevant or redundant attributes, feature selection techniques can significantly reduce the complexity of the original problem, with important benefits in terms of domain understanding and knowledge discovery. When learning from biomedical data, however, the dimensionality issue is often addressed without a joint consideration of other critical aspects that may compromise the performance of the induced models. The adverse implications of an imbalanced class distribution, for example, are often neglected in this domain. The aim of this work is to investigate the effectiveness of hybrid learning strategies that incorporate both methods for dimensionality reduction as well as methods for alleviating the issue of class imbalance. Specifically, we combine different feature selection techniques, both univariate and multivariate, with sampling-based class balancing methods and cost-sensitive classification. The performance of the resulting learning schemes is experimentally evaluated on six high-dimensional genomic benchmarks, using different classification algorithms, with interesting insight about the best strategies to use based on the characteristics of the data at hand.

INDEX TERMS Bioinformatics, class imbalance, cost-sensitive classification, feature selection, high-dimensional data analysis, random forest, random under-sampling, SMOTE over-sampling.

I. INTRODUCTION

Extracting useful knowledge from biomedical datasets is recognized to be a very demanding task. Modern high-throughput technologies, such as mass-spectrometry, DNA micro-arrays and RNA sequencing, have indeed produced an ever-increasing amount of data in recent years, posing unique challenges for the machine learning and data mining communities. A special attention has been given to the automatic classification of cancerous samples based on suitable models built from these kinds of datasets [1].

In this domain, the first and most critical issue is often the huge dimensionality, i.e. the presence of a very high number of attributes (or features) for each of the problem instances at hand [2], [3]. This may negatively impact on the performance of machine learning algorithms, not only in terms of computational efficiency but also in terms of final predictive accuracy, since the generalization ability of the induced models may significantly degrade when the size of the search space is very large (the so-called “curse of dimensionality”

issue) [4]. A proper reduction of the data dimensionality is then of paramount importance, as recognized by a great body of scientific literature in the field [5]–[7].

The available approaches for dimensionality reduction can be broadly categorized into two main groups: (i) *mapping* techniques, that leverage algebraic methods to define new attributes, as combinations of the original ones, and project the data into a lower-dimensional space [8], and (ii) *feature selection* techniques, that attempt to identify the most informative attributes for the task at hand (e.g., based on their correlation with the target class), discarding those that are either irrelevant or redundant [9]. This last approach, which is often preferable in terms of domain understanding and interpretability of the final model, is extensively employed in biomedical data analysis, with an ever-increasing number of reported applications [10]–[13].

Based on their interaction with the algorithm used for inducing the model, feature selection methods are usually distinguished into [6]:

(i) *Filters*, that carry out the selection process as a pre-processing step, without interacting with the classifier; they rely on the intrinsic characteristics of the training data, e.g.,

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Callico^{ID}.

by measuring (through some statistical or entropic criterion) the degree of correlation between the features and the target class [14]. This approach is not tied to a specific learning algorithm and generally has the advantage of a lower computational cost.

(ii) *Wrappers*, that compare different feature subsets and select the one that optimizes the performance of a given classifier. This involves using a suitable search strategy (e.g., a greedy search or an evolutionary search) to build the candidate subsets and evaluating each of them by training and testing a classification model. Tuned to a specific learner, this approach may lead to better results, but at an increased computational cost [5].

(iii) *Embedded methods*, that rely on the intrinsic capacity of some learning algorithms to assess the relevance of the features (*Support Vector Machine* classifiers, for example, allow to weight the features based on the contribution they give to the induced decision function [15]).

Though a lot of research has been devoted to investigating the strengths and weaknesses of the different selection approaches in the biomedical field [4], [5], [10], [14], [16], [17], the choice of the most appropriate method for a given task is often difficult. Due to their computational efficiency, filter methods have been the most used so far, but there is a growing tendency to incorporate them into more sophisticated selection strategies. Indeed, *hybrid* methods, that leverage different heuristics at different stages of the selection process (e.g., reducing the data dimensionality by a filter and then further refining the search by a wrapper) [18], or *ensemble* approaches, that properly combine the outcome of different selectors, are increasingly being explored [19], [20].

However, the high dimensionality is not the only challenge of biomedical data analysis. Another important issue that may worsen the performance of machine learning algorithms, but is often neglected in this domain, is the *imbalance* in the class distribution. This occurs when the data contain quite different numbers of instances for the different classes of interest, which is a rather common situation, for example, in cancer prediction tasks. Traditional classification algorithms may not perform adequately in this scenario, primarily because they are designed to maximize the overall prediction accuracy, with a bias towards the majority class, without regard to the significance of the different classes. As a result, they may exhibit poor performance on the minority class, which is however, in most cases, the class of greatest interest (e.g., due to the vital importance of correctly diagnosing a rare disease). Despite its undoubted relevance for practical applications, the class imbalance problem has not received the attention it would deserve in this domain [21], with a vast majority of literature that focuses on the curse of dimensionality alone, as recently discussed in [22].

Class imbalance, in turn, has been mostly treated in the literature as an independent problem, especially in application fields where the number of features is not so critical. A number of approaches, such as sampling-based balancing methods and cost-sensitive classification, have been proposed

to deal with imbalanced datasets [23], [24], with successful applications in different application contexts, such as anomaly detection [25], fraud detection [26] and fault prediction [27].

In more detail, sampling-based methods modify the distribution of the instances so that the minority class is adequately represented in the dataset used to induce the model. Common sampling-based methods are *random under-sampling*, where instances of the majority class(es) are randomly removed from the dataset, and *random over-sampling*, where new instances of the minority class are artificially created from the existing ones (e.g., by simply duplicating some instances chosen at random). A popular type of over-sampling, known as *SMOTE* (*Synthetic Minority Over-sampling Technique*), involves the generation of new minority instances by interpolating between existing minority instances that are close to each other [28]. This approach is now considered as a de facto standard in the context of learning from imbalanced datasets [29].

On the other hand, cost-sensitive approaches do not modify the distribution of the instances but assign different misclassification costs to the different classes. The underlying ratio is that not all the errors have the same consequences, and practical cost, in real-world applications: misclassifying a rare instance (e.g., a rare disease) is indeed serious. Incorporating misclassification costs into the learning process implies to induce the model with the lowest “overall cost”, instead of the one with the lowest expected error as in the traditional setting.

While several research efforts have explored the issues of high dimensionality and class imbalance independently, only a few studies have addressed both the problems simultaneously [30]–[35]. Since several biomedical datasets are both high-dimensional and class-imbalanced, the aim of this work is to investigate the effectiveness of learning strategies that are designed to handle simultaneously both the issues, in order to effectively deal with real-world problems that involve the classification of rare pathological conditions (e.g., rare cancer types).

Specifically, extending our previous research in this area [22], we explore the combination of sampling-based balancing methods and cost-sensitive classification with suitable feature selection strategies, chosen to be representatives of different selection approaches (both univariate and multivariate).

Using as benchmarks six challenging genomic datasets, we experimentally evaluate the extent to which the resulting learning schemes are advantageous compared to the application, as is common practice in this field, of feature selection alone. The results of the experiments give interesting insight into the benefits of taking class imbalance into account when analyzing such kind of datasets, as well as into the best strategies to use in dependence of the specific characteristics of the data at hand (e.g., the number of the available instances and their level of imbalance).

The remainder of this paper is organized as follows. Section II illustrates all the materials and methods involved in our study, including the considered feature selection techniques and their combination with methods for alleviating the class imbalance problem. The main experimental results are summarized in Section III (and additional results are also given as supplementary material). The findings of the study are further discussed in section IV, which also provides a comparison with the literature. Finally, section V outlines the concluding remarks and some possible directions for future research.

II. MATERIALS AND METHODS

A number of learning strategies that incorporate both methods for dimensionality reduction as well as methods for alleviating the class imbalance problem are here presented. Specifically, sub-section II-A introduces the feature selection techniques, while sub-sections II-B and II-C discuss how feature selection can be combined with class balancing methods and cost-sensitive classification respectively. In the remaining part of the section, we present the datasets and the settings of the experiments, as well as the metrics used for performance evaluation.

A. SELECTING A SUBSET OF MEANINGFUL FEATURES

In general, given a dataset with M features, the output of the feature selection process can be expressed in the form of: (a) a *weighting* of the M features, i.e. each feature is weighted based on a suitable relevance criterion which is usually meant to capture the strength of the correlation between the feature and the target class; (b) a *ranking* of the M features, i.e. the features are ordered based on their relevance, from the most important to the least important (obviously, a feature weighting can be easily converted to a feature ranking by sorting the weights assigned to the features); (c) a *subset* of the M features, which can be selected based on a subset evaluation strategy (as in the wrapper approaches) or simply choosing the “best” features from a list of previously weighted/ranked features (in this case a suitable criterion is required to filter the list).

A very common practice in high-dimensional data analysis, when a subset of meaningful features is required for inducing a descriptive/predictive model, is to use a ranking-based selection approach, coupled with a proper threshold (t), i.e. only the first t top-ranked features are selected, as schematized in Fig. 1. If needed, the resulting feature subset may be further refined through more sophisticated search strategies that – although infeasible in a very large search space – may be still applied after a first, preliminary, dimensionality reduction [36].

The feature selection process depicted in Fig. 1, denoted hereafter as *FS*, can be implemented using different ranking methods. Specifically, we included in our study five techniques that are representative of quite different heuristics. In particular, we considered three univariate methods (*Symmetrical Uncertainty*, *Gain Ratio* and *Chi Squared*), which

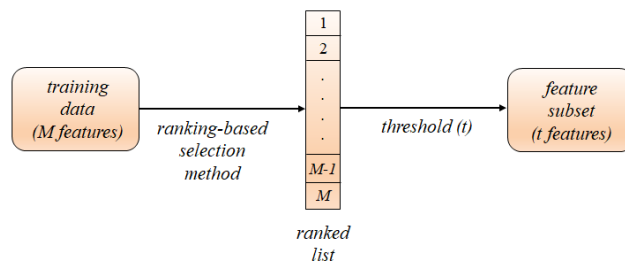


FIGURE 1. Ranking-based feature selection.

assess the relevance of each feature independently from the others, and two multivariate methods (*ReliefF* and *SVM-AW*), which take into account the inter-dependencies among the features. A more detailed description of these techniques, along with a discussion of their pattern of agreement, can be found in [7]. For all of them, we exploited the implementation provided by the WEKA library [37].

In brief:

- *Symmetrical Uncertainty (SU)* and *Gain Ratio (GR)* exploit the concept of *information gain*, which is a measure of the extent to which the class entropy decreases when the value of a given feature is known. However, *SU* and *GR* differ for the way they try to compensate for the information gain’s bias toward features with more values.
- *Chi Squared (χ^2)* evaluates each feature by measuring its chi-squared statistic with respect to the class: the larger the chi-squared, the higher the relevance of the feature for the task at hand.
- *ReliefF* ranks the features based on their ability to differentiate between data instances that are near to each other in the attribute space.
- *SVM-AW* exploits a linear *Support Vector Machine (SVM)* classifier, which has an embedded capability of assigning a weight to each feature (based on the induced hyperplane function [15]); the absolute value of this weight (*AW*) is used to rank the features. The iterative variant of this method, although proposed as a good option for biomedical data analysis [38], is not employed here due to its poor stability [39].

B. USING FEATURE SELECTION IN CONJUNCTION WITH SAMPLING-BASED CLASS BALANCING STRATEGIES

Data sampling is a popular technique used to alleviate the class imbalance problem [24], [40]. As previously mentioned, the basic idea is to modify the proportion between majority and minority instances in the training data. Among the sampling-based approaches, the *random under-sampling*, that involves a reduction of the majority instances, has proved to be effective within different experimental conditions [41]. On the other hand, a duplication of the minority instances (as in the *random over-sampling* approach) may involve a higher risk of overfitting, especially in small sample size settings, which are quite common in the biomedical field. This led us to consider an alternative over-sampling technique, the *SMOTE* approach [28], that has recently been applied in a variety of domains, with quite good results [29].

Specifically, we investigate here the effectiveness of using both *random under-sampling* (hereafter *RUS*) and *SMOTE* in conjunction with feature selection (*FS*). Without loss of generality, we consider a binary classification setting (indeed, a multiclass problem can be always decomposed into a set of binary problems). The evaluation is performed in a two-fold way:

- *Sampling + FS* learning schemes : *RUS(R:1) + FS* and *SMOTE(R:1) + FS*. We first resample the original data by reducing the level of class imbalance to a pre-specified ratio $R:1$, i.e. R instances of the majority class for each instance of the minority class (e.g., $R = 1$ to obtain a uniform class distribution). This is achieved by removing instances of the majority class (*RUS*) or by adding a proper number of synthetic instances of the minority class (*SMOTE*). Feature selection is then performed on the sampled data and, at the end, a classifier is built.

- *FS + Sampling* learning schemes : *FS + RUS(R:1)* and *FS + SMOTE(R:1)*. We first select a subset of meaningful features from the original dataset and then perform data sampling (again, with a pre-specified imbalance ratio $R:1$); as a final step, the classifier is built.

This two-fold setting allows us to investigate the extent to which the final performance is affected by the order of the pre-processing operations performed before constructing the model. Also, the influence of the R parameter (that determines the imbalance ratio in the sample) is experimentally investigated, as detailed in sub-section II-E.

C. USING FEATURE SELECTION IN CONJUNCTION WITH COST-SENSITIVE LEARNING

Differently from traditional classifiers, which try to minimize the overall number of classification errors, cost-sensitive learners attempt to induce the least costly model, provided that different costs are assigned to the different types of errors [23].

As usual practice, we denote here the minority class as positive (+) and the majority class as negative (-). With this notation, the different classification costs can be encoded in a *cost matrix* as the one reported in Table 1. This matrix reflects the fact that we are interested in a model that achieves the best possible performance on the minority class. Indeed, the cost of misclassifying a positive instance (i.e. a false negative error) is C times greater than the cost of misclassifying a negative instance (i.e. a false positive error).

But it is important to remark that, given the high dimensionality of the data here considered, the model is trained after reducing the data dimensionality, i.e. after applying a proper

feature selection technique, so that the learning process can jointly exploit the information conveyed by the selected features as well as the cost information. The effect of varying the C parameter, i.e. the cost of misclassifying a minority instance, is in turn investigated in the experimental study, as explained in sub-section II-E.

D. METRICS FOR PERFORMANCE EVALUATION

The overall percentage of correct predictions, namely the *accuracy*, is the most employed performance measure in the context of classification tasks. It is not meaningful, however, when dealing with quite imbalanced class distributions. In presence of a rare class, indeed, a trivial model that assigns every object to the majority class will have a high level of accuracy even if it fails to recognize any of the rare instances.

More appropriate metrics should then be used to capture the ability of the model to perform well on each single class. Among the evaluation measures proposed in the literature [42], we consider those reported in Table 2, which have proved to be useful in the context of imbalanced classification problems [43].

In the table, the following standard notation is used: TP is the number of positive instances correctly classified (*true positives*); TN is the number of negative instances correctly classified (*true negatives*); FP is the number of negative instances incorrectly classified as positive (*false positives*); FN is the number of positive instances incorrectly classified as negative (*false negatives*).

Note that *specificity* (or TN_rate), *sensitivity* (or *recall* or TP_rate) and *precision* (or *positive predictive value*) incorporate information about only one type of error (false positive or false negative). Hence, it is useful to combine them into an overall evaluation criterion, such as the *F-measure* (harmonic mean between sensitivity and precision) or the *G-mean* (geometric mean between sensitivity and specificity), that accounts for both false positives and false negatives. As well, the *Matthews Correlation (MC) coefficient* takes into account the balance ratios of the four confusion matrix categories (TP , TN , FP , FN), and it is considered an informative score to establish the quality of a binary classifier, even in a class imbalanced scenario [44], [45].

The use of several metrics allows the polyhedral aspects of the classification performance to be captured from different points of views. Of course, metrics different from those here considered could also be used [24], [42], and the best choice of the evaluation framework in the context of imbalanced learning tasks is still a matter of debate [46]. Nevertheless, recent literature confirms the suitability of the approach chosen for this study [43].

E. DATASETS AND EXPERIMENTAL SETTINGS

To evaluate the effectiveness of the learning strategies described in sub-sections II-B and II-C, we performed extensive experiments on six high-dimensional genomic benchmarks [47]–[50], whose main characteristics are summarized in Table 3.

TABLE 1. Cost matrix for a binary classification task.

		predicted class	
		+	-
actual class	+	0	C
	-	1	0

TABLE 2. Performance measures.

Measure	Definition
Specificity or TN_rate (fraction of negative instances classified correctly)	$\frac{TN}{TN + FP}$
Sensitivity or Recall or TP_rate (fraction of positive instances classified correctly)	$\frac{TP}{TP + FN}$
Precision or Positive Predictive Value (fraction of instances that are actually positive in the group the classifier has predicted as positive)	$\frac{TP}{TP + FP}$
F-measure or F₁ score (harmonic mean between sensitivity and precision)	$\frac{2 \cdot \text{sensitivity} \cdot \text{precision}}{\text{sensitivity} + \text{precision}}$
G-mean (geometric mean between sensitivity and specificity)	$\sqrt{\text{sensitivity} \cdot \text{specificity}}$
Matthews Correlation (MC) coefficient (correlation between the observed and predicted classifications)	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

TABLE 3. Genomic datasets used in the experiments.

Dataset	Task	No. of features	No. of instances	% of minority instances (<i>min_pct</i>)
<i>NO-glioma</i>	Discriminating non-classic oligodendroglioma (<i>NO</i>) among a group of glioma instances [47]	12625	50	30%
<i>Lymphoma</i>	Discriminating between follicular lymphoma (<i>FL</i>) and diffuse large b-cell lymphoma (<i>DLBCL</i>) [48]	7129	77	25%
<i>CO-glioma</i>	Discriminating classic oligodendroglioma (<i>CO</i>) among a group of glioma instances [47]	12625	50	14%
<i>Uterus</i>	Discriminating uterus cancer from other cancer types [49]	10935	1545	8%
<i>Omentum</i>	Discriminating omentum cancer from other cancer types [49]	10935	1545	5%
<i>SCLC</i>	Discriminating between small cell lung cancer (<i>SCLC</i>) and different lung tumors [50]	12601	203	3%

All the above benchmarks, which derive from DNA micro-array experiments, contain biological samples described by the expression level of thousands of genes. For each dataset, the features were first ranked using different methods (*SU*, *GR*, χ^2 , *ReliefF*, *SVM-AW*), as described in sub-section II-A. Only the first $t = 100$ top-ranked features, corresponding to about 1% (or less) of the original dimensionality, were retained from the resulting ranked lists. Actually, the optimal value for t may depend on the specific dataset as well as on the applied learning scheme. However, a number of preliminary experiments have shown that a subset size $t = 100$ is a reasonable option across the different datasets and the different settings here considered, while further increasing the number of selected genes does not result in an improved performance. This is in line with the vast literature on micro-array data analysis [5] which highlights the importance of drastically reducing the number of genes to decrease the risk of overfitting and improve the performance (and the understandability) of the induced models.

As regards the classification method, we performed experiments with different learning algorithms: *Random Forest (RF)* [51], *AdaBoost (AdaB)* [52], *k-Nearest Neighbors (k-NN)* [53], *Support Vector Machines (SVM)* [54] and *RIPPER (RIP)* [55], that were chosen as representatives of different classification approaches. For each of them, as for the feature selection methods, we relied on the implementation provided by WEKA [37]. In most cases, we maintained the default parametrization, which has proved to be reliable across several tasks [56]. In particular, for the *RF* classifier, we used 100 trees and $\log_2(t) + 1$ random features, which is a suitable setting for imbalanced problems [57]. For the *AdaB* classifier, built using a decision tree as base learner, we performed 10 iterations (indeed, more iterations did not significantly improve the performance). For the *SVM*, a linear kernel was employed, as common practice in micro-array data analysis [58]. For *k-NN*, the k parameter was set to 5, as lower values may be more sensitive to noise, and the neighbors were weighted by their distance, as in similar studies [21].

TABLE 4. NO-glioma dataset ($\text{min_pct} = 30\%$): model performance using the RF classifier and the SU selection method.

	baseline	FS alone	RUS (1:1) + FS	FS + RUS (1:1)	SMOTE (1:1) + FS	FS + SMOTE (1:1)	FS + Cost (C = 2)	FS + Cost (C = 3)
Specificity	0.91	0.93	0.79	0.86	0.92	0.92	0.84	0.74
Sensitivity	0.57	0.65	0.90	0.84	0.73	0.71	0.83	0.91
Precision	0.67	0.79	0.71	0.77	0.82	0.80	0.74	0.65
F-measure	0.60	0.68	0.77	0.78	0.74	0.72	0.75	0.73
G-mean	0.65	0.73	0.83	0.84	0.79	0.78	0.82	0.81
MC coeff.	0.51	0.62	0.68	0.70	0.68	0.65	0.66	0.62

TABLE 5. Lymphoma dataset ($\text{min_pct} = 25\%$): model performance using the RF classifier and the SU selection method.

	baseline	FS alone	RUS (1:1) + FS	FS + RUS (1:1)	SMOTE (1:1) + FS	FS + SMOTE (1:1)	FS + Cost (C = 2)	FS + Cost (C = 3)
Specificity	0.99	0.95	0.85	0.89	0.95	0.94	0.92	0.87
Sensitivity	0.39	0.73	0.94	0.92	0.78	0.79	0.91	0.94
Precision	0.75	0.85	0.73	0.76	0.87	0.85	0.82	0.74
F-measure	0.49	0.76	0.81	0.82	0.80	0.80	0.84	0.81
G-mean	0.54	0.81	0.89	0.90	0.85	0.85	0.91	0.90
MC coeff.	0.48	0.72	0.75	0.77	0.76	0.75	0.80	0.76

As concerns the specific strategies used to alleviate class imbalance, we explored the following values for the R and C parameters (see sub-sections II-B and II-C), that were used, respectively, to control the imbalance ratio and the cost of misclassifying rare instances: (i) $R = 1$, $R = 2$, $R = 3$; (ii) $C = 2$, $C = 3$, $C = 4$, $C = 5$ (indeed, considering different values would not be beneficial, as shown by the experimental results).

As evaluation protocol, we used a stratified 5-fold cross-validation, which was repeated 10 times, to reduce any bias due to a specific data partitioning. This means that, for each learning scheme, we trained and tested the classification model 50 times, using each time different partitions of data at training and testing stages. The values of the evaluation metrics (specificity, sensitivity, precision, F -measure, G -mean and MC coefficient) were then averaged across the different runs. Note that all the pre-processing steps (feature ranking and sampling) were performed, at each run, only on the training data, to avoid biasing the testing results.

All the experiments were performed within the WEKA machine learning workbench [37], which provides all the necessary pre-processing and data manipulation functions, besides the classification methods.

III. RESULTS

The main results of the experiments are here summarized. Since the RF classifier has proved to be overall the most effective, we start by detailing its performance across the different datasets and the different learning schemes presented in section II; next, a comparative view among the different classifiers will be provided.

As a first point, it is interesting to consider the baseline performance of RF on the original datasets, without any form

of manipulation or dimensionality reduction. The resulting accuracy values, averaged across the 50 training-testing runs, are 81.0%, 83.9%, 87.4%, 93.2%, 95.0% and 97.1%, for *NO-glioma*, *Lymphoma*, *CO-glioma*, *Uterus*, *Omentum* and *SCLC* datasets respectively.

However, as discussed in sub-section II-D, the accuracy simply gives the overall percentage of correctly classified instances, without reflecting the ability of the classifier to discriminate the minority class (which is often the most interesting one, as in this context).

Actually, among the minority (i.e. positive) instances, the average rate of correct prediction (sensitivity) is quite low: it ranges from 0.57 in the *NO-glioma* dataset ($\text{min_pct} = 30\%$), which is only moderately imbalanced, to 0.00 in the most imbalanced datasets, i.e. *Omentum* ($\text{min_pct} = 5\%$) and *SCLC* ($\text{min_pct} = 3\%$), where the baseline classifier assigns all the instances to the majority (i.e. negative) class.

A proper reduction of the data dimensionality improves this baseline performance to a significant extent, as shown in Tables 4-9, where the evaluation results for the six datasets (*NO-glioma*, *Lymphoma*, *CO-glioma*, *Uterus*, *Omentum* and *SCLC* respectively) are summarized in terms of the different metrics reported in Table 2. Specifically, the results in Tables 4-9 refer to the model performance achieved with the SU selection method; the complete results obtained with the other ranking approaches, i.e. GR , χ^2 , $ReliefF$ and $SVM-AW$, are here omitted (for the sake of space and readability) but are made available as supplementary material.

When looking at the second column (' FS alone') of the tables, we can observe indeed the effectiveness of feature selection, as expressed by the values of the F -measure (that gives a trade-off between sensitivity and precision), the G -mean (that gives a trade-off between sensitivity and

TABLE 6. CO-glioma dataset (min_pct = 14%): model performance using the RF classifier and the SU selection method.

	baseline	FS alone	RUS (1:1) + FS	FS + RUS (1:1)	SMOTE (1:1) + FS	FS + SMOTE (1:1)	FS + Cost (C = 2)	FS + Cost (C = 3)
Specificity	1.00	0.98	0.80	0.93	0.97	0.97	0.93	0.89
Sensitivity	0.10	0.49	0.90	0.75	0.62	0.62	0.78	0.87
Precision	0.14	0.54	0.53	0.65	0.65	0.65	0.66	0.64
F-measure	0.11	0.49	0.63	0.66	0.61	0.61	0.67	0.70
G-mean	0.12	0.52	0.81	0.75	0.66	0.66	0.79	0.84
MC coeff.	0.11	0.48	0.59	0.64	0.59	0.60	0.65	0.68

TABLE 7. Uterus dataset (min_pct = 8%): model performance using the RF classifier and the SU selection method.

	baseline	FS alone	RUS (2:1) + FS	FS + RUS (2:1)	SMOTE (2:1) + FS	FS + SMOTE (2:1)	FS + Cost (C = 3)	FS + Cost (C = 4)
Specificity	1.00	0.98	0.93	0.93	0.95	0.95	0.94	0.93
Sensitivity	0.17	0.41	0.81	0.79	0.71	0.72	0.77	0.81
Precision	0.87	0.67	0.49	0.50	0.56	0.57	0.53	0.50
F-measure	0.28	0.50	0.61	0.61	0.63	0.64	0.63	0.62
G-mean	0.41	0.63	0.86	0.86	0.82	0.83	0.85	0.87
MC coeff.	0.36	0.49	0.59	0.59	0.60	0.61	0.60	0.60

TABLE 8. Omentum dataset (min_pct = 5%): model performance using the RF classifier and the SU selection method.

	baseline	FS alone	RUS (2:1) + FS	FS + RUS (2:1)	SMOTE (2:1) + FS	FS + SMOTE (2:1)	FS + Cost (C = 3)	FS + Cost (C = 4)
Specificity	1.00	0.99	0.92	0.93	0.96	0.97	0.96	0.95
Sensitivity	0.00	0.21	0.80	0.80	0.64	0.64	0.66	0.75
Precision	0.00	0.58	0.36	0.37	0.48	0.51	0.50	0.46
F-measure	0.00	0.30	0.49	0.50	0.54	0.56	0.57	0.56
G-mean	0.00	0.45	0.86	0.86	0.78	0.78	0.80	0.84
MC coeff.	0.00	0.32	0.50	0.51	0.52	0.54	0.55	0.55

TABLE 9. SCLC dataset (min_pct = 3%): model performance using the RF classifier and the SU selection method.

	baseline	FS alone	RUS (2:1) + FS	FS + RUS (2:1)	SMOTE (2:1) + FS	FS + SMOTE (2:1)	FS + Cost (C = 3)	FS + Cost (C = 4)
Specificity	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
Sensitivity	0.00	0.50	0.77	0.95	0.58	0.58	0.99	1.00
Precision	0.00	0.56	0.47	0.95	0.64	0.64	0.97	0.93
F-measure	0.00	0.52	0.54	0.95	0.60	0.60	0.97	0.96
G-mean	0.00	0.52	0.78	0.95	0.61	0.61	0.99	1.00
MC coeff.	0.00	0.52	0.56	0.95	0.60	0.60	0.98	0.96

specificity) and the *MC* coefficient (that captures the degree of correlation between actual classification and model prediction).

Nevertheless, these results are still sub-optimal, and the main aim of this study is to investigate the extent to which a further improvement is achievable, in terms of final predictive performance, by properly combining feature selection with methods that cope with the class imbalance problem. Specifically, as discussed in sub-sections II-B and II-C, we experimented with hybrid strategies where the feature selection (*FS*) process is carried out:

- after data sampling (*RUS + FS* and *SMOTE + FS* learning schemes)
- before data sampling (*FS + RUS* and *FS + SMOTE* learning schemes)
- in conjunction with cost-sensitive classification (*FS + Cost* learning scheme)

All the sampling-based learning schemes have been implemented with different imbalanced ratios *R*:1 (i.e. *R* negative instances for each positive instance); in particular, only the settings *R* = 1 and *R* = 2 have been considered for the *NO-glioma* and *Lymphoma* datasets, which are moderately

imbalanced, while the setting $R = 3$ has also been explored for the other benchmarks. However, for the sake of space, only the results achieved with $R = 1$ are reported for the first three benchmarks (Tables 4-6), while those achieved with $R = 2$ are given for the most imbalanced datasets (Tables 7-9). Indeed, these settings have proved to be suitable. As well, for the $FS + Cost$ scheme, different values of the C parameter (cost of misclassifying a positive instance) have been considered, as explained in sub-section II-E, but only the most interesting results are summarized in the Tables 4-9.

As we can see, when compared to the application of feature selection alone, the sampling-based learning schemes ($RUS + FS$, $FS + RUS$, $SMOTE + FS$, $FS + SMOTE$), as well as the cost-sensitive scheme ($FS + Cost$), result in a further increased sensitivity (that reflects the ability of the model to detect the class of interest, i.e. the minority class). When RUS sampling and costs are used, the improvement in sensitivity is more pronounced but, in most cases, at the expenses of lower values of specificity and precision. On the other hand, the increase in sensitivity is less pronounced with $SMOTE$, which however does not impact negatively on specificity and precision (with a few exceptions). Despite the above differences, both the “trade-off” measures, i.e. the F -measure and the G -mean, show that the hybrid learning strategies, which cope simultaneously with both high-dimensionality and class imbalance, are generally better than using feature selection alone, especially when the percentage of minority instances is quite low. This is also confirmed by the MC coefficient.

Note that the statistical significance of the observed differences has been first evaluated according to a paired t -test [56], at a confidence level of 0.05: the performance values that turned out to be significantly better than ‘ FS alone’ are marked in bold in the tables. A further discussion of the statistical significance of the findings of our study, using a more restrictive version of the standard t -test, will be given in the next section.

To facilitate the comparison among the different learning schemes and help the reader to assess their effectiveness in dependence on the level of class imbalance, Figs. 2 and 3 show a histogram representation of the last three metrics (F -measure, G -mean and MC coefficient) for the six benchmarks here considered. Specifically, Fig. 2 refers to the *NO-glioma*, *Lymphoma* and *CO-glioma* datasets, while Fig. 3 refers to the *Uterus*, *Omentum* and *SCLC* datasets, that are the most imbalanced ones; in both these figures, as well as in those shown later, standard error bars are also included.

Noteworthy, the results shown so far (which refer to the performance achieved with the SU selection method) are in great part consistent with the results obtained with the other selection methods (i.e. GR , χ^2 , *ReliefF*, *SVM-AW*), attached as supplementary material. As a representative example, a comparative view is given in Fig. 4 for the *Omentum* dataset ($min_pct = 5\%$), which has shown to be a particularly challenging benchmark. As we can see, the univariate selection methods (SU , GR and χ^2) lead to very similar results, both when used alone as well as in conjunction with RUS ,

SMOTE and cost-sensitive learning. Among the multivariate selection methods, *ReliefF* performs somewhat better, while *SVM-AW* leads to the worst results when used alone but greatly improves its performance in conjunction with strategies that alleviate the class imbalance problem (leading, in the hybrid setting, to results comparable with those achieved with the other selection methods).

Finally, it is interesting to extend the analysis to the other classification algorithms considered in this study (*AdaB*, *k-NN*, *SVM* and *RIP*). As anticipated, the best results have been achieved with the *RF* classifier, which is increasingly being used across several domains [59]–[62]. Specifically, when both the high-dimensionality and the class imbalance are properly addressed, *RF* outperforms the other methods in terms of F -measure and MC coefficient, as shown in Fig. 5 for the *Omentum* benchmark. Nevertheless, it is worth of remark that all the considered classifiers benefit from the application of the hybrid learning strategies here discussed, which lead to better results compared to the application of feature selection alone. The only exception is the *RIP* algorithm, which is already designed to cope with imbalanced class distributions [55], thus benefiting to a lesser extent from the use of strategies that further tackle class imbalance (*SMOTE*, however, is still somewhat useful in this case too).

A deeper discussion of the findings of this study and a comparison with the literature will be given in the next section.

IV. DISCUSSION

As shown previously, a wide experimental study has been conducted on six high-dimensional genomic benchmarks which present different levels of class imbalance, with a percentage of minority instances that ranges from 30% to 3% (Table 3). The study has encompassed different classifiers, different feature selection methods and different learning strategies, giving useful insight along the following dimensions:

(i) *Need for handling class imbalance.* Despite several biomedical datasets present imbalanced class distributions, most of the studies in the field have so far focused on the dimensionality issue alone. Indeed, a proper reduction of the data dimensionality is of paramount importance in terms of domain understanding as well as for obtaining better predictive models, as confirmed by the results shown in section III. However, when feature selection is combined with methods to tackle class imbalance, such as resampling and cost-sensitive learning, the classification performance can be further improved to a significant extent. The superiority of the hybrid learning strategies that address both high dimensionality and class imbalance has been shown in the previous section using the widely employed paired t -test, at a confidence level of 0.05. The corrected resampled t -test [56], recently proposed as a more reliable option for repetitive random sampling and cross-validation experiments, confirms that the hybrid strategies are significantly better than feature selection alone for the most imbalanced datasets (*Uterus*, *Omentum* and *SCLC*). Hence, in case of highly skewed

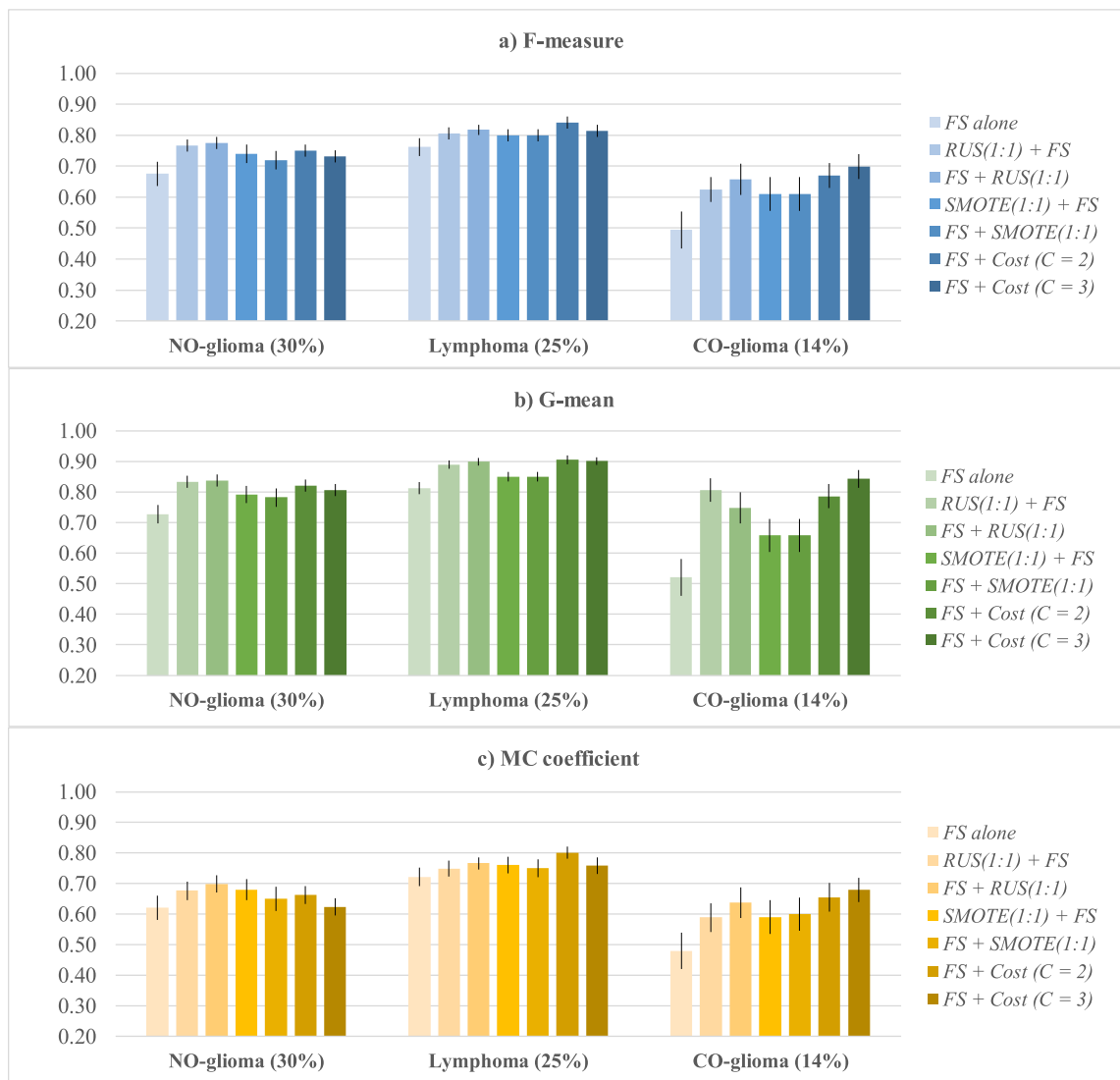


FIGURE 2. Classification performance, in terms of a) F-measure, b) G-mean and c) MC coefficient, for the NO-glioma, Lymphoma and CO-glioma datasets, using the RF classifier and the SU selection method, in conjunction with different learning strategies.

distributions, the adoption of proper strategies that cope with class imbalance is a primary need, since feature selection alone may not be enough to achieve satisfactory results. On the other hand, handling class imbalance alone, without reducing the data dimensionality, is not an option in this domain since the selection of a small number of features is crucial in terms of knowledge discovery (e.g., to better understand the genetic basis of cancer). Further, a preliminary ablation study has shown that sampling-based class balancing strategies and cost-sensitive learning are overall less effective when used alone, without feature selection. For example, in the two datasets with the lowest percentage of minority instances, i.e. *Omentum* and *SCLC*, the RF classifier achieves an MC coefficient of 0.41 and 0.42 respectively, if used with a cost $C = 3$ alone, without any dimensionality reduction, whereas the corresponding values in conjunction with the SU selector ($FS + Cost$ scheme) are higher, as shown in the

Tables 8-9. As well, using the $SMOTE(2:1)$ approach alone, RF gives an MC coefficient of 0.45 and 0.20 for *Omentum* and *SCLC* respectively, while the corresponding values incorporating the SU selector (either before or after sampling) are better (again, see the Tables 8-9). Both the F-measure and the G-mean values confirm this trend.

(ii) *Choice of the hybrid learning scheme.* As highlighted previously, combining feature selection with a proper strategy to tackle class imbalance turns out to be more effective, in the specific domain here considered, than using feature selection alone or class balancing strategies alone. All the hybrid learning schemes investigated in this study ($RUS + FS$, $FS + RUS$, $SMOTE + FS$, $FS + SMOTE$, $FS + Cost$) have proved to be somewhat useful, and there is no scheme that is always better than the others, as shown in section III. However, using $SMOTE$ seems to be less convenient when the absolute number of positive instances in the training data

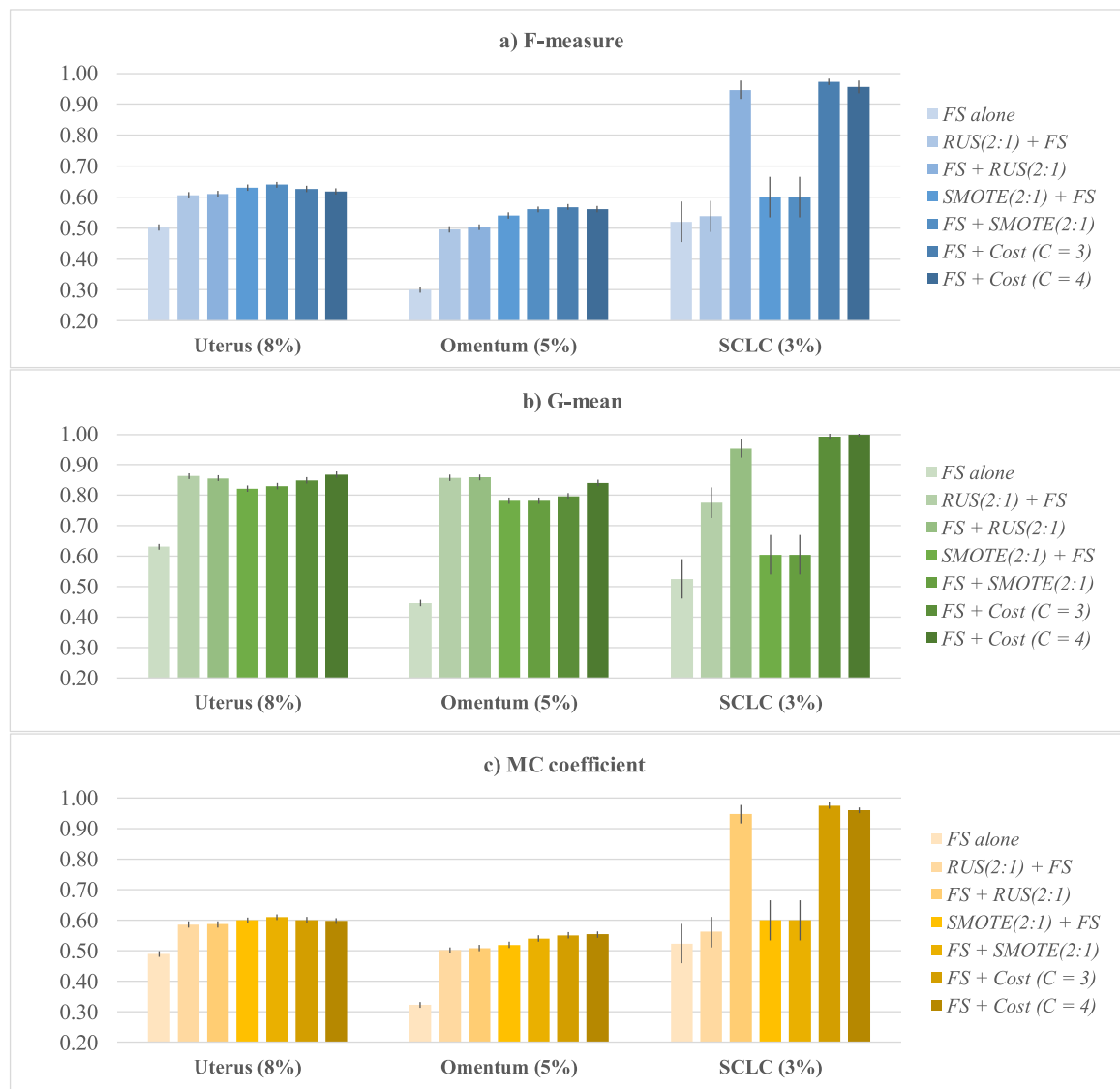


FIGURE 3. Classification performance, in terms of a) F-measure, b) G-mean and c) MC coefficient, for the Uterus, Omentum and SCLC datasets, using the RF classifier and the SU selection method, in conjunction with different learning strategies.

is particularly low, as in the *CO-glioma* and *SCLC* datasets. In turn, when using *RUS*, performing feature selection after data sampling does not turn out to be advantageous in highly imbalanced and small sample size settings, as in the *SCLC* dataset. For both *RUS* and *SMOTE*, the imbalance ratio in the resampled data should be chosen in dependence on the original percentage of minority instances. Making the class distribution uniform has indeed turned out to be a good option only when the original percentage of minority instances is higher than 10%; instead, in case of more skewed datasets, the settings *RUS(2:1)* and *SMOTE(2:1)* have proved to be better. As well, when a cost-sensitive approach is adopted, the cost of misclassifying a minority instance (expressed by the C parameter in our methodology) should be increased in dependence of the level of class imbalance in the original dataset (anyway, values above $C = 4$ do not seem to be beneficial).

(iii) *Choice of the methods for classification and feature selection.* As shown in section III, different classification algorithms may benefit from the adoption of a hybrid learning scheme. In particular, among the classifiers here considered, *RF* is the one that achieves the best performance when both the high dimensionality and the class imbalance are properly addressed (Fig. 5). This confirms that the *RF* classifier, already known to be effective both on biomedical data [59] as well as in different application fields [60]–[62], can be successfully applied also in presence of imbalanced class distributions [57], [63]. On the other hand, there is no selection method that clearly outperforms the others. Indeed, as shown in Fig. 4 (and, more comprehensively, in the attached supplementary material), the different selection methods often lead to comparable results, with a few exceptions especially for the multivariate approaches (*SVM-AW*, in particular, performs worse in some cases). But it is worth to remark that, regardless

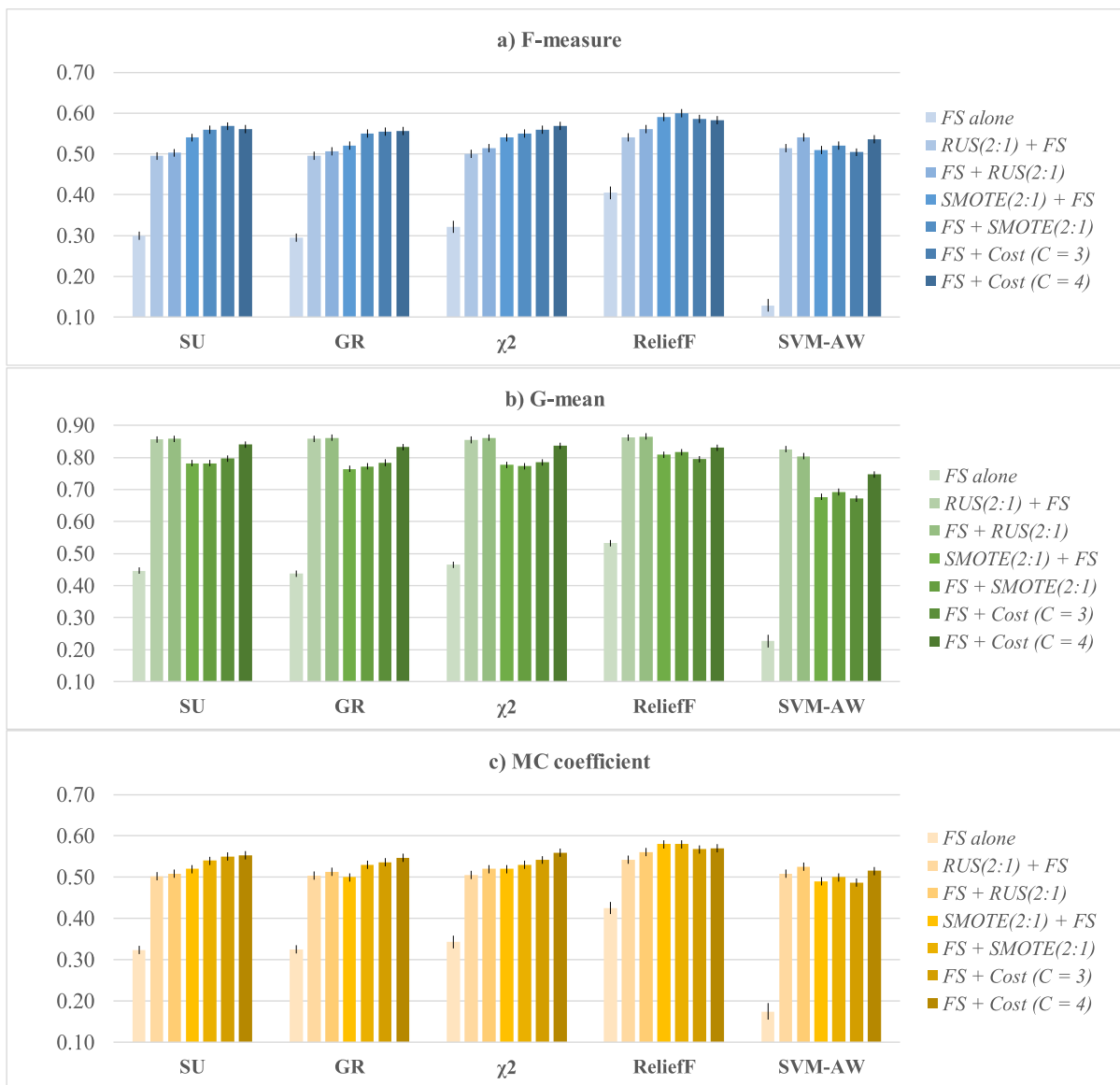


FIGURE 4. Omentum dataset ($min_pct = 5\%$): classification performance, in terms of a) F-measure, b) G-mean and c) MC coefficient, achieved with SU, GR, χ^2 , ReliefF and SVM-AW selection methods, using the RF classifier, in conjunction with different learning strategies.

of the specific selection approach, the benefits of carrying out feature selection increase when the class imbalance problem is properly addressed.

The findings of this work are partially consistent with the results recently discussed in [21], where the effectiveness of combining RUS and feature selection is evaluated in conjunction with different classifiers and feature selection methods, but within less severe imbalance settings ($min_pct > 10\%$). The beneficial impact of sampling-based approaches on high-dimensional bioinformatics datasets is also explored in [30]–[32]. In particular, [30] relies on both RUS and feature selection, and investigates the extent to which the order of these pre-processing operations impacts on the classification results. As well, [31] exploits both RUS and feature selection and shows that using fully balanced data significantly

improves the SVM performance in protein function prediction tasks. An evaluation of SMOTE for high-dimensional class-imbalanced micro-array data is presented in [32], where only k -NN classifiers are found to take significant advantage of SMOTE over-sampling, provided that the number of features is properly reduced (differently from our study, however, all the experiments are conducted on datasets with $min_pct \geq 14\%$).

A discussion of the issues associated with class prediction in high dimensional domains can also be found in [33], where both simulated and real genomic data are used to investigate the effectiveness of a RUS-based ensemble strategy (where different bootstrap samples are built by removing instances of the majority class), in conjunction with specific combinations of classifiers and feature selection methods.

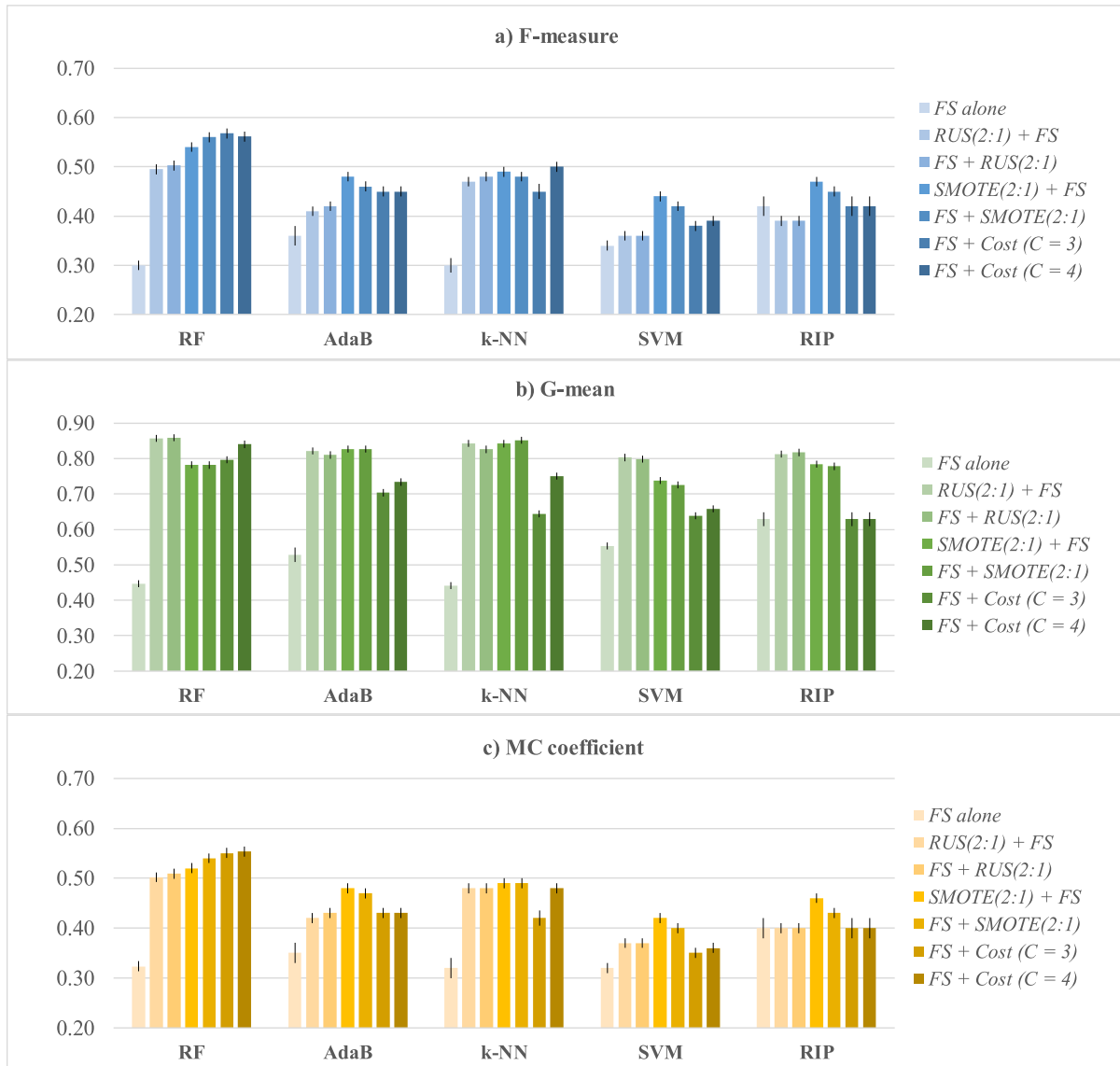


FIGURE 5. Omentum dataset (min_pct = 5%): classification performance, in terms of a) F-measure, b) G-mean and c) MC coefficient, achieved with RF, AdaB, k-NN, SVM and RIP classifiers, using the SU selection method, in conjunction with different learning strategies.

A different ensemble strategy, which relies on a random subspace approach, is presented in [34]: though the developed solution shows good performance on imbalanced data, and seems also to be robust to dimension increase, it has not been evaluated on benchmarks with thousands of features, as the ones here considered. Finally, ad hoc improvements of existing classification algorithms have been proposed, e.g. in [35], to cope with problems that are both high-dimensional and class-imbalanced.

Despite the valuable works mentioned above, many key research issues remain to be addressed in this domain, with a lack of comparative studies which may provide insight on which approach would be most appropriate in a given scenario. Our analysis, although not exhaustive, is an attempt to provide a contribution in this direction. In particular, compared to other studies that address similar tasks, this work pro-

vides a comparative evaluation which encompasses a wider range of techniques (*RUS + FS*, *FS + RUS*, *SMOTE + FS*, *FS + SMOTE*, *FS + Cost*) and a wider range of class imbalance levels (from 30% to 3% of minority instances).

V. CONCLUDING REMARKS

Using six challenging genomic benchmarks, this study has evaluated the effectiveness of hybrid learning strategies that try to cope with high-dimensional and class-imbalanced data. Specifically, we have explored different ways of combining feature selection with sampling-based balancing methods (*random under-sampling* and *SMOTE*) and cost-sensitive learning.

Encompassing different levels of class imbalance, as well as different classification algorithms (*RF*, *AdaB*, *k-NN*, *SVM* and *RIP*) and selection methods (*SU*, *GR*, χ^2 , *ReliefF* and

SVM-AW), our study has shown that the explored hybrid strategies are overall more effective than using feature selection alone, especially when the class distribution is highly skewed. Some insight has also been gained on which strategies, and parameter settings, may be more convenient based on the characteristics of the data at hand.

To further strengthen the conclusions of the study and better understand the combined effects of class imbalance and high-dimensionality in the biomedical domain, we plan to extend our experiments along multiple directions. First, a larger number of datasets with different characteristics will be analyzed. As well, other feature selection approaches, besides the ranking-based methods so far considered, will be evaluated in conjunction with methods for handling class imbalance. In particular, it could be interesting to explore the extent to which emerging feature selection paradigms, such as ensemble feature selection and deep learning feature selection, may be beneficial in this type of application. Finally, a multi-faceted evaluation will be performed using different evaluation metrics, in order to comprehensively characterize the impact of the considered learning strategies.

REFERENCES

- [1] A. K. Tanwani, J. Afridi, M. Z. Shafiq, and M. Farooq, "Guidelines to select machine learning scheme for classification of biomedical datasets," in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* (Lecture Notes in Computer Science), vol. 5483. Berlin, Germany: Springer, 2009, pp. 128–139.
- [2] R. Clarke, H. W. Resson, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, "The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data," *Nature Rev. Cancer*, vol. 8, no. 1, pp. 37–49, Jan. 2008.
- [3] V. Bolón-Canedo, N. Sánchez-Maróño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Inf. Sci.*, vol. 282, pp. 111–135, Oct. 2014.
- [4] P. Drotár, J. Gazda, and Z. Smékal, "An experimental comparison of feature selection methods on two-class biomedical datasets," *Comput. Biol. Med.*, vol. 66, pp. 1–10, Nov. 2015.
- [5] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007.
- [6] V. Bolón-Canedo, N. Sánchez-Maróño, and A. Alonso-Betanzos, *Feature Selection for High-Dimensional Data* (Artificial Intelligence: Foundations, Theory, and Algorithms). Cham, Switzerland: Springer, 2015.
- [7] N. Dessì and B. Pes, "Similarity of feature selection methods: An empirical study across data intensive classification tasks," *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4632–4642, Jun. 2015.
- [8] J. J. Dai, L. Lieu, and D. Rocke, "Dimension reduction for classification with gene expression microarray data," *Stat. Appl. Genet. Mol. Biol.*, vol. 5, no. 1, 2006, Art. no. 6.
- [9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
- [10] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Adv. Bioinf.*, vol. 2015, pp. 1–13, 2015, Art. no. 198363.
- [11] V. Bolón-Canedo, N. Sánchez-Maróño, and A. Alonso-Betanzos, "Distributed feature selection: An application to microarray data classification," *Appl. Soft Comput.*, vol. 30, pp. 136–150, May 2015.
- [12] B. Pes, "Evaluating feature selection robustness on high-dimensional data," in *Proc. 13th Int. Conf. Hybrid Artif. Intell. Syst. (HAIS)*, in Lecture Notes in Artificial Intelligence, vol. 10870. Cham, Switzerland: Springer, 2018, pp. 235–247.
- [13] D. Zhang, L. Zou, X. Zhou, and F. He, "Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer," *IEEE Access*, vol. 6, pp. 28936–28944, 2018.
- [14] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. De Schaezzen, R. Duque, H. Bersini, and A. Nowe, "A Survey on filter techniques for feature selection in gene expression microarray analysis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 4, pp. 1106–1119, Jul. 2012.
- [15] A. Rakotomamonjy, "Variable selection using SVM based criteria," *J. Mach. Learn. Res.*, vol. 3, pp. 1357–1370, Mar. 2003.
- [16] A.-C. Hauray, P. Gestraud, and J.-P. Vert, "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures," *PLoS ONE*, vol. 6, no. 12, Dec. 2011, Art. no. e28210.
- [17] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," in *Proc. IEEE 13th Int. Conf. Inf. Reuse Integr. (IRI)*, Aug. 2012, pp. 356–363.
- [18] N. Almgren and H. Alshamlan, "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification," *IEEE Access*, vol. 7, pp. 78533–78548, 2019.
- [19] N. Dessì and B. Pes, "Stability in biomarker discovery: Does ensemble feature selection really help?" in *Proc. 28th Int. Conf. Ind., Eng. Other Appl. Appl. Intell. Syst. (IEAAIE)*, in Lecture Notes in Computer Science, vol. 9101. Cham, Switzerland: Springer, 2015, pp. 191–200.
- [20] A. Ben Brahim and M. Limam, "Ensemble feature selection for high dimensional data: A new method and a comparative study," *Adv. Data Anal. Classification*, vol. 12, no. 4, pp. 937–952, Dec. 2018.
- [21] A. Abu Shanab and T. Khoshgoftaar, "Is gene selection enough for imbalanced bioinformatics data?" in *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2018, pp. 346–355.
- [22] B. Pes, "Handling class imbalance in high-dimensional biomedical datasets," in *Proc. IEEE 28th Int. Conf. Enabling Technol., Infrastruct. Collaborative Enterprises (WETICE)*, Jun. 2019, pp. 150–155.
- [23] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009.
- [24] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.
- [25] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, "Toward credible evaluation of anomaly-based intrusion-detection methods," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 5, pp. 516–524, Sep. 2010.
- [26] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," *J. Netw. Comput. Appl.*, vol. 68, pp. 90–113, Jun. 2016.
- [27] Z. Wu, Y. Guo, W. Lin, S. Yu, and Y. Ji, "A weighted deep representation learning model for imbalanced fault diagnosis in cyber-physical systems," *Sensors*, vol. 18, no. 4, p. 1096, Apr. 2018.
- [28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jul. 2018.
- [29] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, Jul. 2018.
- [30] T. M. Khoshgoftaar, A. Fazelpour, D. J. Dittman, and A. Napolitano, "Classification performance of three approaches for combining data sampling and gene selection on bioinformatics data," in *Proc. IEEE 15th Int. Conf. Inf. Reuse Integr. (IRI)*, Aug. 2014, pp. 315–321.
- [31] A. Al-Shahib, R. Breitling, and D. Gilbert, "Feature selection and the class imbalance problem in predicting protein function from sequence," *Appl. Bioinf.*, vol. 4, no. 3, pp. 195–203, 2005.
- [32] R. Blagus and L. Lusa, "Evaluation of SMOTE for high-dimensional class-imbalanced microarray data," in *Proc. 11th Int. Conf. Mach. Learn. Appl.*, Dec. 2012, pp. 89–94.
- [33] W.-J. Lin and J. J. Chen, "Class-imbalanced classifiers for high-dimensional data," *Briefings Bioinf.*, vol. 14, no. 1, pp. 13–26, Jan. 2013.
- [34] P. Ksieniewicz and M. Woźniak, "Dealing with the task of imbalanced, multidimensional data classification using ensembles of expositors," in *Proc. Mach. Learn. Res.*, vol. 74, 2017, pp. 164–175.
- [35] R. Blagus and L. Lusa, "Improved shrunken centroid classifiers for high-dimensional class-imbalanced data," *BMC Bioinf.*, vol. 14, Dec. 2013, Art. no. 64.
- [36] L. M. Cannas, N. Dessì, and B. Pes, "A filter-based evolutionary approach for selecting features in high-dimensional micro-array data," in *Proc. 6th Int. Conf. Intell. Inf. Process. (IIP)*, 2010, pp. 297–307.
- [37] *Weka 3: Data Mining Software in Java*. [Online]. Available: <https://www.cs.waikato.ac.nz/ml/weka/>

- [38] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.
- [39] B. Pes, "Feature selection for high-dimensional data: The issue of stability," in *Proc. IEEE 26th Int. Conf. Enabling Technol., Infrastruct. Collaborative Enterprises (WETICE)*, Jun. 2017, pp. 170–175.
- [40] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. Int. Conf. Artif. Intell.*, Halifax, NS, Canada, Jun. 2000, pp. 111–117.
- [41] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 935–942.
- [42] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Jun. 2009.
- [43] A. Luque, A. Carrasco, A. Martín, and A. De Las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, Jul. 2019.
- [44] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData Mining*, vol. 10, no. 1, p. 35, 2017, doi: [10.1186/s13040-017-0155-3](https://doi.org/10.1186/s13040-017-0155-3).
- [45] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric," *PLoS ONE*, vol. 12, no. 6, Jun. 2017, Art. no. e0177678.
- [46] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–50, Aug. 2016.
- [47] C. L. Nutt, D. R. Mani, R. A. Betensky, and P. Tamayo, "Gene expression-based classification of malignant gliomas correlates better with survival than histological classification," *Cancer Res.*, vol. 63, no. 7, pp. 1602–1607, 2003.
- [48] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Med.*, vol. 8, no. 1, pp. 68–74, Jan. 2002.
- [49] *OpenML Datasets*. Accessed: Nov. 2019. [Online]. Available: <https://www.openml.org/search?type=data>
- [50] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 24, pp. 13790–13795, Nov. 2001.
- [51] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [52] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Mach. Learn.*, 1996, pp. 148–156.
- [53] R. M. Parry, W. Jones, T. H. Stokes, J. H. Phan, R. A. Moffitt, H. Fang, L. Shi, A. Oberthuer, M. Fischer, W. Tong, and M. D. Wang, "K-nearest neighbor models for microarray gene expression analysis and clinical outcome prediction," *Pharmacogenomics J*, vol. 10, no. 4, pp. 292–309, Aug. 2010.
- [54] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," *Neural Comput.*, vol. 13, no. 3, pp. 637–649, Mar. 2001.
- [55] W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn.*, 1995, pp. 115–123.
- [56] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.
- [57] T. M. Khoshgoftaar, M. Golawala, and J. V. Hulse, "An empirical study of learning from imbalanced data using random forest," in *Proc. 19th IEEE Int. Conf. Tools Artif. Intell. (ICTAI)*, Oct. 2007, pp. 310–317.
- [58] S. Mukherjee, "Classifying microarray data using support vector machines," in *A Practical Approach to Microarray Data Analysis*. Boston, MA, USA: Springer, 2003.
- [59] N. Dessì, G. Milia, and B. Pes, "Enhancing random forests performance in microarray data classification," in *Proc. 14th Conf. Artif. Intell. Med. (AIMS)*, in Lecture Notes in Computer Science, vol. 7885. Springer, 2013, pp. 99–103.
- [60] L. Rokach, "Decision forest: Twenty years of research," *Inf. Fusion*, vol. 27, pp. 111–125, Jan. 2016.
- [61] O. Bălan, G. Moise, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, "Fear level classification based on emotional dimensions and machine learning techniques," *Sensors*, vol. 19, no. 7, p. 1738, Apr. 2019.
- [62] J. Ma, Y. Qiao, G. Hu, Y. Huang, A. K. Sangaiah, C. Zhang, Y. Wang, and R. Zhang, "De-anonymizing social networks with random forest classifier," *IEEE Access*, vol. 6, pp. 10139–10150, 2018.
- [63] K. Toyoda, P. Takis Mathiopoulou, and T. Ohtsuki, "A novel methodology for HYIP operators' bitcoin addresses identification," *IEEE Access*, vol. 7, pp. 74835–74848, 2019.



BARBARA PES (Member, IEEE) was born in Cagliari, Italy, in 1976. She received the Laurea degree in physics from the University of Cagliari, in 2001.

From 2002 to 2005, she collaborated with the Database and Data Mining Group, Department of Mathematics and Computer Science, University of Cagliari. Since 2006, she has been with the Department of Mathematics and Computer Science, as a University Researcher (permanent position). Here she teaches/taught Foundations of Computer Science, Database, and Data Mining Courses. She has participated in several research projects on web-based information systems, service-oriented architectures, data integration, high-dimensional data analysis, and bio-informatics. Her main research interests are in the field of data mining and machine learning, classification of high-dimensional data, and feature selection. She is the author of more than 70 articles published in international conferences, books, and journals.

• • •