IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 17, 2022, accepted June 1, 2022, date of publication June 8, 2022, date of current version June 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3180773

Multivariate Feature Ranking With High-Dimensional Data for Classification Tasks

FERNANDO JIMÉNEZ[1], GRACIA SÁNCHEZ[1], JOSÉ PALMA[1], (Senior Member, IEEE), LUIS MIRALLES-PECHUÁN[2], AND JUAN A. BOTÍA[1]

[1]Department of Information and Communication Engineering, University of Murcia, 30071 Murcia, Spain
[2]School of Computer Science, Technological University Dublin, Dublin 7, D07 EWV4 Ireland

Corresponding author: Fernando Jiménez (fernan@um.es)

This work was supported in part by the Clinical Decision Support System for Infection Surveillance (SITSUS) Project by the Spanish Ministry of Science, Innovation and Universities (MCIU) under Grant RTI2018-094832-B-I00; in part by the Spanish Agency for Research (AEI); in part by the European Fund for Regional Development (FEDER); and in part by the Science and Technology Agency, Séneca Foundation, Comunidad Autónoma Región de Murcia, Spain, under Project 00004/COVI/20 and Project 00007/COVI/20.

ABSTRACT In many machine learning classification problems, datasets are usually of high dimensionality and therefore require efficient and effective methods for identifying the relative importance of their attributes, eliminating the redundant and irrelevant ones. Due to the huge size of the search space of the possible solutions, the attribute subset evaluation feature selection methods are not very suitable, so in these scenarios feature ranking methods are used. Most of the feature ranking methods described in the literature are univariate methods, which do not detect interactions between factors. In this paper, we propose two new multivariate feature ranking methods based on pairwise correlation and pairwise consistency, which have been applied for cancer gene expression and genotype-tissue expression classification tasks using public datasets. We statistically proved that the proposed methods outperform the state-of-the-art feature ranking methods *Clustering Variation*, *Chi Squared*, *Correlation*, *Information Gain*, *ReliefF* and *Significance*, as well as other feature selection methods for attribute subset evaluation based on correlation and consistency with the multi-objective evolutionary search strategy, and with the embedded feature selection methods *C4.5* and *LASSO*. The proposed methods have been implemented on the WEKA platform for public use, making all the results reported in this paper repeatable and replicable.

INDEX TERMS High-dimensional data, classification, feature ranking, feature selection, machine learning, correlation, consistency.

I. INTRODUCTION

High-dimensional classification is one of the main machine learning tasks that has been addressed in the literature during the last decade [1]–[3]. The elimination of redundant and irrelevant attributes through the application of *feature selection* (FS) methods, although demanding a huge search space, allows reducing the complexity of the classification models while improving their accuracy. One of the classification problems with high-dimensional data that has had the greatest impact on the scientific community is the *gene expression* (GE) classification, particularly for cancer classification. GE problems currently represent an excellent testbed for experimentation and comparison of FS techniques in classification tasks with high-dimensional data. That is why in this paper, GE is used as a reference framework. Filter-based FS has been widely used in the literature with GE data. The main disadvantage is that filter techniques only take into account characteristics inherent to the data, regardless of the task to be solved (diagnosis, prognosis, or clustering). The use of filter-based FS techniques has increased the range of performance measures to be used for classification tasks, allowing their integration with search strategies for subset evaluation. Filter methods have also been used successfully for *feature ranking* (FR). FR methods assign a ranking or importance to each attribute, and they can also be treated as FS techniques if a subset with the q best attributes in the ranking, or those above a certain threshold t of importance, is selected

The associate editor coordinating the review of this manuscript and approving it for publication was Victor S. Sheng.

(see Fig. 1). Examples of filter methods applied to GE data for classification tasks include mutual information [4], information gain [5], minimum redundancy - maximum relevance [6] and symmetric uncertainty [7]. An exhaustive analysis of filter techniques for FS in GE data can be found in [8].

Another type of FS methods used for high-dimensional classification problems are the wrapper methods. In contrast to a filter-based method, wrapper FS methods build a predictive model to evaluate attributes, either individually or attribute subsets. Obviously, wrapper FS methods require more computational time than filter-based methods since a predictive model has to be fitted for each candidate subset. The main advantage of this approach relies on the fact that attribute evaluation is task-oriented. In other words, attributes are evaluated according to their predictive power by improving the predictive model performance. For example, in [9] support vector machines have been used with recursive feature elimination as the search strategy whilst in [10] was used a best-first search. Apart from their high computation cost, another disadvantage of wrapper methods is that predictive model overfitting can affect the solution quality. To overcome this problem, there is a growing interest in the use of hybrid techniques, which try to integrate two or more different FS methods, finding informative attributes and reducing the computational cost [11]. Moreover, different subsets search strategies have been proposed to improve the efficiency of FS methods applied to GE data, such as metaheuristic techniques [12]. These strategies do not guarantee to find the optimal subsets but reach acceptable solutions in terms of a trade-off between optimality and computing effort. Their main advantage is that due to their global optimization approach, they avoid being trapped in a local minimum (or local maximum) as it can be the case for deterministic techniques. Additionally, the metaheuristic techniques allow defining more than one objective and several constraints to the optimization problem.

A third group of FS methods used on high dimensional datasets are the embedded methods. In these methods, the FS algorithm is integrated as part of the learning algorithm. In [13], an FS algorithm for high-dimensional microarray data is proposed, which first uses a mutual information method to filter out irrelevant genes, and then uses an improved LASSO-based method [14] to remove redundant genes. In [15], a combination of filter (ReliefF [16]), wrapper (with greedy stepwise search [17]) and embedded (LASSO) methods is used for gene expression data. In [18], an embedded strategy that penalizes the cardinality of the feature set via the scaling factors technique is proposed, and is used with support vector machines (SVM) on highly imbalanced microarray datasets.

Since high-dimensional classification problems contain thousands of attributes, subset evaluation FS methods are often inefficient in this scenario. The search space of these FS problems is $O(2^n)$, where n is the number of attributes, and heuristics and metaheuristics can only find satisfactory solutions using very long computation times, even if the FS method is of filter type. The computation time required increases, even more, when we use wrapper feature selection methods, which can become impractical. FR methods, which evaluate attributes individually instead of evaluating attribute subsets, are probably the most viable alternative for this type of scenario. However, existing FR methods are univariate methods, except *ReliefF* which is a multivariate FR method. Univariate FR methods evaluate the attributes in a ''myopic'' way. That is, not considering interdependencies or interactions between the attributes. If each attribute is evaluated without considering the rest of the attributes, we will probably obtain poor results, because the attribute interdependencies are not being considered. In this paper, we propose two novel multivariate FR methods based on pairwise correlation and pairwise consistency respectively. The proposed FR methods are compared with a wide range of FR methods (including the multivariate *ReliefF* method), with subset evaluation FS methods based on correlation and consistency with a multi-objective evolutionary search strategy, and with the embedded feature selection methods *C4.5* and *LASSO*. We have used a GE cancer RNA-Seq dataset to perform the experiments given the current importance of this type of application, and two additional genotype-tissue expression datasets (brain and age) to confirm the results.

The paper has been organized as follows: section II shows the state-of-the-art of FS methods for high-dimensional data, with special attention to GE classification; section III describes the novel multivariate FR methods proposed in this paper; section IV describes the datasets and the performed experiments; section V analyses the obtained results; finally, section VI presents conclusions and outlines for future work.

II. RELATED WORKS

A FS wrapper method of subset evaluation where the search strategy is a competitive swarm optimizer (CSO) is proposed in [19]. CSO is a recent variant of PSO which has been dedicated to large-scale optimization, hence the authors use it to solve high-dimensional FS problems. Since CSO was originally developed for continuous optimization, it must be adapted to combinatorial optimization to solve FS problems. The authors use the k-nearest-neighbour (kNN) classifier to test the effectiveness of the FS. An archive technique is also used to reduce the computational cost. In [20], the authors propose a new approach based on iteratively adjusting a bound on the l_1-norm of a SVM in order to force the number of selected features to converge towards the desired maximum limit. In [21], a real-coded genetic algorithm is used to optimize attribute weights that minimize the average of prediction errors of the entire training dataset of a weighted kNN classifier. The proposed method is evaluated on six high-dimensional microarray datasets. In [22], an ensemble-based FS method that combines random bits forest and recursive clustering elimination is proposed. The authors also introduce an FS stability measurement method, which measures whether the FS is stable or not through the intersection measurement. In [23], a comparison of the

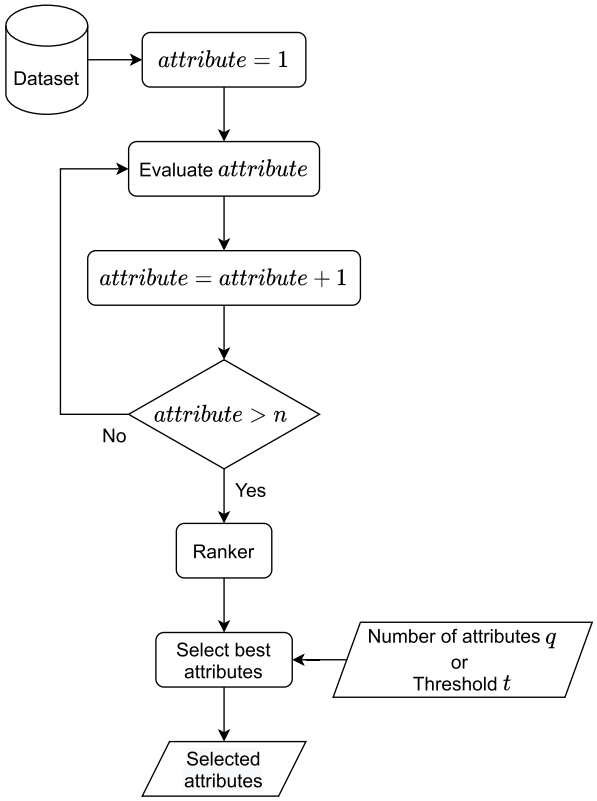

FIGURE 1. Feature selection based on feature ranking.

single-objective and multi-objective approaches for FS and classification is performed. In the single-objective approach, analysis of variance (ANOVA) and Chi-Square have been used. The features selected have been used for building twelve classifications to find out which combination presents better performance. In the multi-objective approach, a metaheuristic wrapper based technique is used to simultaneously find the best combination of feature subsets and classification techniques. In this case, the metaheuristic techniques used are the NSGA-II multi-objective evolutionary algorithm and the Multi-dimensional Archive of Phenotypic Elites (MAP-Elites) algorithm. The results reveal that the two approaches produce good outcomes, and the computational cost of the models based on FR techniques is substantially lower. However, metaheuristics based approaches perform a better exploration of the search spaces. Specifically, NSGA-II showed a tendency in selecting fewer features whereas MAP-Elites presented a wider Pareto front achieving competitive results.

An interesting benchmark of ranker methods in high-dimensional GE applied to eleven survival datasets is presented in [24]. Specifically, fourteen ranker based filter methods have been applied. Among them, we can find those based on variance, correlation, Cox score and mutual information (MI). Ranker based filter methods are compared with regularized Cox proportional hazards models using all features. The drawn conclusions state that the models obtained after applying one of the filters achieve better predictive power than the baseline model. Furthermore, models with better predictive power only select a small number of features. One of the most interesting characteristics of ranker methods is that they provided, as output, a ranking of the features based on a concrete univariate measure. This facilitates the aggregation of different FR methods as part of an ensemble-based strategy. An example of this approach is presented in [25]. In this work, different versions of the microarray data set are generated by bootstrapping the original set. In each bootstrapped bag a different ranker technique is applied. The final ranking is the result of the average ranking of each feature across all the partial rankings. This approach has also been successfully applied in [26] for stomach cancer biomarker identification. In this case, four different ranking techniques have been applied: Conditional Mutual Information Maximisation (CMIM) [27], Double Input Symmetrical Relevance (DISR) [28], Interaction Capping (ICAP) [29] and Conditional Informative Feature Extraction (CIFE) [30]. A novel technique, Weighted Ensemble of Ranks (WER_j) is proposed to aggregate individual rankings. Then, among the top 100 genes, only those common to all different models are selected. The proposed method suggests that the selected genes show better performance accuracy when multiple clinical outcomes are considered.

Despite MI being one of the most used measures for FR, obtaining a reliable estimation of MI on high-dimensional low-sample datasets remains a challenge. In this regard, in [31], a novel Joint Bias Mutual Information (JBMI) is first presented, together with modified Discretization and Selection of features based on MI (mDSM). The authors also demonstrate that MI follows a χ^2 distribution, making it possible to design an FS technique that, simultaneously with FR, selects the best discretization of selected features using the χ^2 criteria. These findings have been used in [32], where a novel Mutual information-based Gene Selection (MGS) is presented. In order not to lose relevant genes, mDSM is applied in a Leave-One-Out Cross-Validation scheme, resulting in different rankings of relevant genes. Then two ranking criteria have been applied to aggregate different rankings namely MGS frequency-based ranking (MGS_f) and MGS Random Forest based ranking (MGS_{rf}). An evaluation of the proposed techniques has been performed over different GE datasets. Results proved that both (MGS_f) and (MGS_{rf}) outperforms existing techniques in both balanced and imbalanced datasets.

An interesting approach using rankers based on Autoencoders (AEs) is presented in [33]. First, an AE is applied for a non-linear fusion and summarization of the original characteristics. Then, the FR technique ANOVA with FDR correction is used to retain the most relevant features that have been used in different machine learning (ML) classification techniques. The conclusions showed that the combination of AE and FR provides better performance. It could be argued that the use of AE does not allow the extraction of biomedical related conclusions since some information about the

original features was lost in the AE fusion process. However, it is worth mentioning that a methodology to calculate gene weights for genes of a set of AE features is also presented.

III. MATERIALS AND METHODS

Let $D = \{I_1, \ldots, I_w\}$ be a dataset with w instances. Each instance $I_p = (a_1^p, \ldots, a_n^p, c_p)$, $p = 1 \ldots, w$, has n input attributes of any type, and one output attribute $c_p \in \{1, \ldots, s\}$, where s is the number of output classes. We assume that at least one instance exists for each output class. In the following lines, we describe two new multivariate FR methods for high-dimensional data. The first method is based on pairwise correlation and the second one on pairwise consistency.

A. MULTIVARIATE FEATURE RANKING BASED ON PAIRWISE CORRELATION

We propose the FR method called *Pairwise Correlation*,[1] which is inspired in the *correlation-based feature selection* (CFS) method. The CFS algorithm was developed by Mark A. Hall from the University of Waikato in Hamilton (New Zealand) throughout his doctoral thesis [34]. This same university is well-known in the world of data science for developing the free software application WEKA (*Waikato Environment for Knowledge Analysis* [35]), in which Mark A. Hall has an important role as Honorary Research Associate. The CFS algorithm has the advantage over other FS methods that it generates more accurate models and reduces the number of attributes selected by half in most cases. This concept is explained years later in a summary paper [36]. The CFS method evaluates sets of attributes instead of doing so individually. To determine the goodness of each set, the CFS algorithm evaluates how well each attribute can predict the class as well as the similarity degree between the attributes. In such a way, the feature sets correlated with the class and with features poorly correlated with each other obtain the higher scores. CFS method uses the function $\Phi_D(\mathcal{S})$ to measure the quality of a subset $\mathcal{S}$ of k attributes in a dataset D, $1 \leq k \leq n$, defined as follows:

$$\Phi_D(\mathcal{S}) = \frac{k \cdot \sigma_D^c}{\sqrt{k + k \cdot (k-1) \cdot \sigma_D^f}} \tag{1}$$

where σ_D^c is the mean of the correlations between each feature in $\mathcal{S}$ and the class attribute, and σ_D^f is the average correlation between each of the $\binom{k}{2}$ possible feature pairs in $\mathcal{S}$. In other words, the numerator indicates the predictive degree of a set of variables while the denominator indicates the redundancy between the variables. CFS method requires discretizing the values (usually with Fayyad and Irani method [37]). CFS applies the *symmetrical uncertainty* method [38] to measure the degree of similarity for discrete values.

[1]The *Pairwise Correlation* method has been incorporated into the WEKA platform as an official package with the name *PairwiseCorrelationAttributeEval*.

The proposed FR method *Pairwise Correlation* evaluates an attribute $i \in \{1, \ldots, n\}$ by using the following function Φ_D^A:

$$\Phi_D^A(i) = \sum_{\substack{j \in \{1,\ldots,n\} \\ j \neq i}} \Phi_D(\{i, j\}) \tag{2}$$

where $\Phi_D(\{i, j\})$ is the merit (eq. (1)) of the subset formed by attributes i and j, for all $j = 1, \ldots, n$, with $j \neq i$. That is, the merit $\Phi_D^A(i)$ of an attribute i is the sum of the merits Φ_D of the attribute subsets formed by i and each of the other attributes. Attributes with low correlation to other attributes and highly correlated with the class are preferred. *Pairwise Correlation* is a (filter) multivariate FR method since the evaluation of each attribute takes into account all the other attributes together with the class, thus considering the interactions between the attributes.

The computational complexity of the *Pairwise Correlation* method to evaluate n attributes is $O(n^2 \cdot w)$, since $\Phi_D(\{i, j\})$ is computed with complexity $O(w)$, and therefore $\Phi_D^A(i)$ is computed with complexity $O(n \cdot w)$ for each attribute i. Since there are n attributes, the total computational complexity to evaluate all the n attributes is $O(n^2 \cdot w)$.

B. MULTIVARIATE FEATURE RANKING BASED ON PAIRWISE CONSISTENCY

Similarly, we propose the FR method called *Pairwise Consistency*,[2] which uses the *consistency* metric for attribute subsets introduced by Liu and Setiono [39]. The intuition behind the consistency measure is to find attributes that divide the dataset into parts with a highly predominant class. This measure has been explained by Almuallim and Dietterich [40] in 1991 and by Liu and Setiono [39] in 1996, but the research of Dash and Liu [41] gives a more complete perspective. According to [41], a group of features is *inconsistent* when two or more instances have the same values but different labels. For example, the instances $(1, 2, 2, a)$ and $(1, 2, 2, b)$, where a and b represent the class, are inconsistent. The consistency measure is then defined by the *inconsistency rate*. The *inconsistency rate*, $I_D(\mathcal{S})$, of an attribute subset $\mathcal{S}$ in a dataset D is calculated as the sum of all the inconsistency counts for all the patterns divided by the total number of instances in D. The inconsistency count for a given pattern (the values of the selected features without the class) is calculated as the total number of the same patterns in the dataset minus the number of instances of the majority class of the pattern. For example, for the pattern $(1, 2, 2)$, if there are 36 elements of class a, 6 for class b, and 5 for class c, the inconsistency count would be $(36 + 6 + 5) - 36 = 11$. Obviously, the less inconsistent the subset is, the greater the consistency of an attribute subset. In summary, the consistency measure is monotonic, fast, multivariate, able to remove redundant and/or irrelevant features, and capable of handling

[2]The *Pairwise Consistency* method has been incorporated into the WEKA platform as an official package with the name *PairwiseConsistencyAttributeEval*.

some noise [41]. The consistency of any subset can never be lower than that of the full set of attributes. The usual practice is to use this subset evaluator in conjunction with a search strategy that looks for the smallest subset with consistency equal to that of the full set of attributes. The consistency measure can work when data has discrete-valued features. Any continuous feature should be first discretized using some discretization method [37].

The proposed FR method *Pairwise Consistency* evaluates an attribute $i \in \{1, \ldots, n\}$ by using the following function Ψ_D^A:

$$\Psi_D^A(i) = \sum_{\substack{j \in \{1,\ldots,n\} \\ j \neq i}} \Psi_D(\{i, j\}) \tag{3}$$

where $\Psi_D(\{i, j\}) = 1 - I_D(\{i, j\})$ is the *consistency rate* of the subset formed by the attributes i and j, for all $j = 1, \ldots, n$, with $j \neq i$. That is, the merit $\Psi_D^A(i)$ of an attribute i is the sum of the consistency rates of the attribute i and each of the other attributes. *Pairwise Consistency* is also a (filter) multivariate FR method and therefore considers interactions between factors.

The computational complexity of the *Pairwise Consistency* method to evaluate n attributes is $O(n^2 \cdot w)$, since $\Psi_D(\{i, j\})$ is computed with complexity $O(w)$ (using a hashing mechanism as indicated in [41]), and therefore $\Psi_D^A(i)$ is computed with complexity $O(n \cdot w)$ for each attribute i. Since there are n attributes, the total computational complexity to evaluate all the n attributes is $O(n^2 \cdot w)$.

IV. EXPERIMENTS AND RESULTS

In this section, we describe the gene expression datasets used in this paper (sections IV-A and IV-B) as well as the experiments performed and their results (summarized in section IV-C). The *Pairwise Correlation* and *Pairwise Consistency* methods have been compared with 6 FR methods. For this, a set of 8 classification algorithms has been used, and the obtained models have been compared using as a metric the percentage of correct predictions. Statistical tests are performed to detect statistically significant differences between FR methods and to establish a win-loss ranking. Furthermore, the proposed FR methods have been compared with other filter FS methods for attribute subset evaluation that use a metaheuristic search strategy, and with the embedded feature selection methods *C4.5* and *LASSO*. Sections with runtimes, external validation, and t-SNE visualization are also included.

A. GENE EXPRESSION CANCER RNA-SEQ DATASET

The *gene expression cancer RNA-Seq* dataset[3] is part of the RNA-Seq (HiSeq) PANCAN dataset. The original dataset[4] is maintained by the cancer genome atlas pan-cancer analysis project [42]. This dataset is a random extraction of gene expressions of patients having different types of tumours, containing 801 instances and 20,531 attributes. Attributes of each instance are RNA-Seq gene expression levels measured by the Illumina HiSeq platform. A dummy name (*gene_xxxxx*) is given to each attribute. Attributes are ordered consistently with the original submission. It is possible to obtain the complete list of names by accessing the project platform. The output attribute has 5 classes corresponding to the 5 tumour types shown in Table 1 along with the number of instances of each class.

TABLE 1. Classes and their number of instances of the gene expression cancer RNA-Seq dataset.

Tumour	*Number of instances*
BRCA	300
COAD	78
KIRC	146
LUAD	141
PRAD	136

1) COMPARISON WITH OTHER FEATURE RANKING METHODS

Our two methods *Pairwise Correlation* and *Pairwise Consistency* have been compared with 6 other well-known FR methods that use different information measures. These FR methods are *Clustering Variation* [43], *Chi Squared* [44], *Correlation* [45], *Information Gain* [46], *ReliefF* [16] and *Significance* [47]. To compare the FR methods, the reduced datasets containing the q best attributes obtained with each FR method have been used to construct classifiers of different nature, for $q = 3, \log_2(n), 50, 100$. The used classification algorithms were *naive bayes* [48], *multilayer perceptron* [49], *support vector machine* [50], *k-NN* [51], *RIPPER* [52], *C4.5* [53], *random forest* [54] and *zeroR* [55]. Each pair (reduced dataset, classification algorithm) has been evaluated using 10-fold cross-validation, repeated 10 times. Therefore, for each pair (reduced dataset, classification algorithm), 100 classifiers have been built, which have been evaluated with the metric of *per cent correct* and the mean has been calculated. Tables 2 to 5 show these results, for each value of q. Table 6 shows, as a summary, the number of times that each FR method has obtained the best evaluation result for some value of q, and the total number. The FR methods were ordered according to this score and the rank of each FR method is shown. The FR methods in the first two positions in the ranking are marked in bold.

We have performed statistical tests to detect statistically significant differences between the compared FR methods. A *paired t-test* has been performed establishing as baseline the reduced dataset obtained with each of the FR methods, for each value of q. In this way, each FR method has been statistically compared with all the others with each of the classifiers. Then, the times that each FR method has won (wins) and the times that it has lost (losses) are obtained. Table 7 shows the results of the statistical tests for each value of q. In this table, the FR methods with the greatest difference between wins and losses in each value of q have been marked in bold. Finally,

[3]https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

[4]https://www.synapse.org/#!Synapse:syn4301332

TABLE 2. Average performance evaluation with the gene expression cancer RNA-Seq dataset for $q = 3$, 10-fold cross-validation, 10 repetitions.

Method	*NaiveBayes*	*MLP*	*SVM*	*kNN*	*RIPPER*	*C4.5*	*Random Forest*	*ZeroR*
Clustering Variation	22.27	37.15	36.84	36.63	37.10	37.45	36.97	37.45
Chi Squared	91.16	86.50	88.53	91.90	90.53	90.97	**93.40**	37.45
Correlation	87.34	75.57	88.99	87.08	89.23	88.91	88.99	37.45
Information Gain	91.16	86.43	88.53	91.90	90.82	90.94	93.28	37.45
ReliefF	91.30	81.38	87.93	89.44	90.19	90.29	91.88	37.45
Significance	91.16	86.54	88.50	91.90	90.52	90.82	93.31	37.45
Pairwise Correlation	92.57	89.64	91.24	92.64	91.11	91.97	93.36	37.45
Pairwise Consistency	92.66	88.99	90.85	89.93	89.28	91.85	92.47	37.45

TABLE 3. Average performance evaluation with the gene expression cancer RNA-Seq dataset for $q = \log_2(n) = 14$, 10-fold cross-validation, 10 repetitions.

Method	*NaiveBayes*	*MLP*	*SVM*	*kNN*	*RIPPER*	*C4.5*	*Random Forest*	*ZeroR*
Clustering Variation	19.84	37.37	38.08	35.79	37.34	37.45	35.94	37.45
Chi Squared	98.99	93.17	99.16	99.10	95.78	96.47	99.21	37.45
Correlation	97.07	86.82	98.14	97.08	92.61	93.56	97.37	37.45
Information Gain	99.10	92.83	99.48	99.25	95.76	96.12	99.01	37.45
ReliefF	96.73	89.71	97.84	97.50	94.92	95.84	97.99	37.45
Significance	98.43	92.49	99.49	99.25	96.24	96.43	99.10	37.45
Pairwise Correlation	99.08	93.26	99.71	**99.75**	96.24	97.05	99.18	37.45
Pairwise Consistency	99.01	92.48	99.20	99.25	95.18	96.22	99.18	37.45

TABLE 4. Average performance evaluation with the gene expression cancer RNA-Seq dataset for $q = 50$, 10-fold cross-validation, 10 repetitions.

Method	*NaiveBayes*	*MLP*	*SVM*	*kNN*	*RIPPER*	*C4.5*	*Random Forest*	*ZeroR*
Clustering Variation	26.90	42.27	41.66	39.15	43.43	43.13	41.65	37.45
Chi Squared	99.60	97.29	99.84	99.85	97.42	97.69	99.73	37.45
Correlation	98.84	96.41	99.63	99.64	95.96	95.77	99.25	37.45
Information Gain	99.74	97.21	**99.88**	**99.88**	97.03	97.64	99.69	37.45
ReliefF	98.86	98.01	99.73	99.48	97.28	97.99	99.55	37.45
Significance	98.56	96.90	99.61	99.64	96.77	97.72	99.50	37.45
Pairwise Correlation	98.63	97.33	99.86	99.85	97.53	97.40	99.51	37.45
Pairwise Consistency	99.56	97.57	99.75	**99.88**	97.12	97.38	99.64	37.45

TABLE 5. Average performance evaluation with the gene expression cancer RNA-Seq dataset for $q = 100$, 10-fold cross-validation, 10 repetitions.

Method	*NaiveBayes*	*MLP*	*SVM*	*kNN*	*RIPPER*	*C4.5*	*Random Forest*	*ZeroR*
Clustering Variation	40.82	45.40	44.81	41.50	47.07	49.25	48.70	37.45
Chi Squared	99.61	97.29	99.76	99.75	97.67	97.75	99.61	37.45
Correlation	99.50	95.38	99.88	99.85	96.25	96.37	99.54	37.45
Information Gain	99.79	98.06	**100.00**	99.88	97.08	97.43	99.78	37.45
ReliefF	99.36	97.78	99.90	**100.00**	97.59	97.63	99.54	37.45
Significance	99.08	98.10	**100.00**	99.75	97.77	97.29	99.64	37.45
Pairwise Correlation	99.44	98.30	**100.00**	**100.00**	97.53	97.90	99.65	37.45
Pairwise Consistency	99.79	97.81	**100.00**	99.78	97.09	97.59	99.76	37.45

TABLE 6. Number of times each method was the best along with the rank for each method.

Method	*Best evaluations*	*Rank*
Clustering Variation	0	4
Chi Squared	1	3
Correlation	0	4
Information Gain	**3**	**1**
ReliefF	1	3
Significance	1	3
Pairwise Correlation	**3**	**1**
Pairwise Consistency	**2**	**2**

Tables 8 to 11 show the summary of the evaluations and statistical tests for each value of q. An entry of the form 'a (b)' in these tables represents the number 'a' of datasets in which the column has been better than the row, and the number 'b' of datasets in which the column has been statistically better than row.

2) COMPARISON WITH ATTRIBUTE SUBSET EVALUATION FEATURE SELECTION METHODS

This section compares the FR methods *Pairwise Correlation* and *Pairwise Consistency* with FS methods that use the correlation and consistency filters but for attribute subset evaluation instead of attribute evaluation. These FS methods require a strategy to search for candidate subsets of attributes in a search space $O(2^n)$ (see Fig. 2). We have used a multi-objective evolutionary search strategy [56]–[60], in particular, the NSGA-II algorithm [61], with which the merit of the attribute subsets is maximized and its cardinality is minimized. In this paper, these FS methods are called

TABLE 7. Wins − losses ranking tests for gene expression cancer RNA-Seq dataset, 10-fold cross-validation, 10 repetitions.

	$q = 3$			$q = \log_2(n) = 14$			$q = 50$			$q = 100$		
Method	***wins***	***losses***	***dif.***	***wins***	***losses***	***dif.***	***wins***	***losses***	***dif.***	***wins***	***losses***	***dif.***
Clustering Variation	0	49	-49	0	49	-49	0	49	-49	0	49	-49
Chi Squared	12	4	8	17	1	16	10	0	10	7	0	7
Correlation	7	24	-17	7	33	-26	7	7	0	7	1	6
Information Gain	12	4	8	17	0	17	12	0	**12**	8	0	**8**
ReliefF	10	8	2	8	18	-10	8	1	7	7	0	7
Significance	12	4	8	17	0	17	8	3	5	7	2	5
Pairwise Correlation	22	0	**22**	19	0	**19**	8	3	5	8	0	**8**
Pairwise Consistency	19	1	18	17	1	16	10	0	10	8	0	**8**

TABLE 8. Summary of the evaluations and statistical tests for gene expression cancer RNA-Seq dataset, $q = 3$, 10-fold cross-validation, 10 repetitions.

	Clustering Variation	*Chi Squared*	*Correlation*	*Information Gain*	*ReliefF*	*Significance*	*Pairwise Correlation*	*Pairwise Consistency*
Clustering Variation	–	7 (7)	7 (7)	7 (7)	7 (7)	7 (7)	7 (7)	7 (7)
Chi Squared	0 (0)	–	1 (0)	1 (0)	1 (0)	1 (0)	6 (2)	4 (2)
Correlation	0 (0)	6 (4)	–	6 (4)	6 (3)	6 (4)	7 (5)	7 (4)
Information Gain	0 (0)	3 (0)	1 (0)	–	1 (0)	2 (0)	7 (2)	4 (2)
ReliefF	0 (0)	6 (1)	1 (0)	6 (1)	–	6 (1)	7 (3)	6 (2)
Significance	0 (0)	4 (0)	1 (0)	3 (0)	1 (0)	–	7 (2)	4 (2)
Pairwise Correlation	0 (0)	1 (0)	0 (0)	0 (0)	0 (0)	0 (0)	–	1 (0)
Pairwise Consistency	0 (0)	3 (0)	0 (0)	3 (0)	1 (0)	3 (0)	6 (1)	-

TABLE 9. Summary of the evaluations and statistical tests for gene expression cancer RNA-Seq dataset, $q = \log_2(n) = 14$, 10-fold cross-validation, 10 repetitions.

	Clustering Variation	*Chi Squared*	*Correlation*	*Information Gain*	*ReliefF*	*Significance*	*Pairwise Correlation*	*Pairwise Consistency*
Clustering Variation	–	7 (7)	7 (7)	7 (7)	7 (7)	7 (7)	7 (7)	7 (7)
Chi Squared	0 (0)	–	0 (0)	3 (0)	0 (0)	3 (0)	6 (1)	3 (0)
Correlation	0 (0)	7 (6)	–	7 (7)	5 (1)	7 (6)	7 (7)	7 (6)
Information Gain	0 (0)	4 (0)	0 (0)	–	0 (0)	4 (0)	6 (0)	3 (0)
ReliefF	0 (0)	7 (4)	2 (0)	7 (3)	–	7 (4)	7 (3)	7 (4)
Significance	0 (0)	4 (0)	0 (0)	3 (0)	0 (0)	–	7 (0)	3 (0)
Pairwise Correlation	0 (0)	1 (0)	0 (0)	1 (0)	0 (0)	0 (0)	–	1 (0)
Pairwise Consistency	0 (0)	4 (0)	0 (0)	4 (0)	0 (0)	4 (0)	6 (1)	–

TABLE 10. Summary of the evaluations and statistical tests for gene expression cancer RNA-Seq dataset, $q = 50$, 10-fold cross-validation, 10 repetitions.

	Clustering Variation	*Chi Squared*	*Correlation*	*Information Gain*	*ReliefF*	*Significance*	*Pairwise Correlation*	*Pairwise Consistency*
Clustering Variation	–	7 (7)	7 (7)	7 (7)	7 (7)	7 (7)	7 (7)	7 (7)
Chi Squared	0 (0)	–	0 (0)	3 (0)	2 (0)	1 (0)	3 (0)	2 (0)
Correlation	0 (0)	7 (1)	–	7 (2)	6 (1)	4 (1)	6 (1)	7 (1)
Information Gain	0 (0)	4 (0)	0 (0)	–	3 (0)	1 (0)	2 (0)	2 (0)
ReliefF	0 (0)	5 (0)	1 (0)	4 (1)	–	1 (0)	3 (0)	4 (0)
Significance	0 (0)	6 (1)	3 (0)	6 (1)	6 (0)	–	6 (0)	6 (1)
Pairwise Correlation	0 (0)	3 (1)	1 (0)	5 (1)	4 (0)	1 (0)	–	4 (1)
Pairwise Consistency	0 (0)	5 (0)	0 (0)	4 (0)	3 (0)	1 (0)	3 (0)	–

TABLE 11. Summary of the evaluations and statistical tests for gene expression cancer RNA-Seq dataset, $q = 100$, 10-fold cross-validation, 10 repetitions.

	Clustering Variation	*Chi Squared*	*Correlation*	*Information Gain*	*ReliefF*	*Significance*	*Pairwise Correlation*	*Pairwise Consistency*
Clustering Variation	–	7 (7)	7 (7)	7 (7)	7 (7)	7 (7)	7 (7)	7 (7)
Chi Squared	0 (0)	–	2 (0)	5 (0)	3 (0)	4 (0)	5 (0)	5 (0)
Correlation	0 (0)	5 (0)	–	7 (0)	6 (0)	5 (0)	6 (1)	6 (0)
Information Gain	0 (0)	2 (0)	0 (0)	–	3 (0)	2 (0)	4 (0)	2 (0)
ReliefF	0 (0)	4 (0)	1 (0)	4 (0)	–	4 (0)	5 (0)	4 (0)
Significance	0 (0)	2 (0)	2 (0)	4 (1)	3 (0)	–	5 (0)	4 (1)
Pairwise Correlation	0 (0)	2 (0)	1 (0)	2 (0)	1 (0)	1 (0)	–	2 (0)
Pairwise Consistency	0 (0)	2 (0)	1 (0)	3 (0)	3 (0)	2 (0)	4 (0)	–

MOEA-CFS and *MOEA-Consistency*. In the comparisons, we have used $q = 100$ and $q = 200$ for the FR methods *Pairwise Correlation* and *Pairwise Consistency*. Again, each pair (reduced dataset, classification algorithm) has been evaluated using 10-fold cross-validation, repeated 10 times. For the sake of a fair comparison, methods *MOEA-CFS* and *MOEA-Consistency* have been run with a number of evaluations of the objective function such that the runtimes are not shorter

than the runtimes required by the *Pairwise Correlation* and *Pairwise Consistency* methods respectively. Thus, the number of evaluations of the objective function given to method *MOEA-CFS* was 30,000,000 (population size of 100 and 300,000 generations), and that of method *MOEA-Consistency* were 130,000,000 (population size of 100 and 1,300,000 generations). Table 12 shows the average evaluations with the metric of *per cent correct*. Table 13 shows the results of the statistical tests and the wins-losses ranking, marking in bold the top 2 methods.

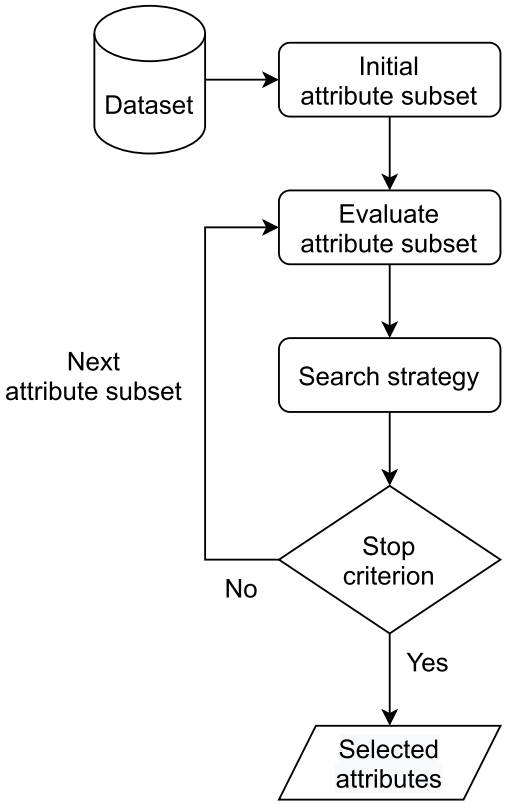

FIGURE 2. Feature selection based on attribute subset evaluation.

3) COMPARISON WITH EMBEDDED FEATURE SELECTION METHODS

Finally, in this section, we show the results of the third group of experiments where the *Pairwise Correlation* and *Pairwise Consistency* methods proposed in this paper are compared with two well-known embedded feature selection methods such as *C4.5* and *LASSO*. We have also included the results of the top three statistically best classifiers with all 20,531 attributes. Table 14 shows the per cent correct in 10-fold cross-validation of the classifiers, as well as the number of used attributes. Internally selected attributes are also shown for embedded methods. The number of attributes and classifiers used for the *Pairwise Correlation* and *Pairwise Consistency* methods correspond to those that have produced the best per cent correct (100%) in the previous experiments shown in sections IV-A1 and IV-A2.

4) RUNTIMES

This section shows the runtime spent by the methods compared in this paper, including the attribute evaluation methods, the attribute subset evaluation methods, and the embedded methods. We have used an Intel (R) Core (TM) i9-10900K CPU 3.70GHz 128 GB RAM x64-based processor, 64-bit operating system. Table 15 shows the runtime, in minutes, of a single execution of each method using the full training set, as well as the ranking of the methods from lowest to highest runtime.

5) EXTERNAL VALIDATION

We wanted to perform a partial external validation of the approach based on answering the following biological question on the PANCAN cancer dataset:

> ''are the proposed feature selection methods effective on selecting the genes that help identify samples of BRCA cancer?''

We propose a discovery-replication approach to provide the answers. For each method *Pairwise Correlation* and *Pairwise Consistency* we present a three-phase strategy, namely (1) discovery, (2) replication and (3) statistical assessment. The first phase is concerned with selecting the genes that best identify BRCA subjects. Therefore, at discovery, we apply feature selection on BRCA and KIRC samples. Then, the features selected in the discovery phase are assessed for replication by testing how well they discriminate between BRCA and each of the other cancer types. The feature selection method should work properly when the classifier built at the discovery phase shows high accuracy. Besides, we should not be able to observe any statistically significant differences between that classifiers accuracy and the accuracy of any of the classifiers built with the same genes and the samples of BRCA and the other cancer types. More specifically, at discovery, we perform feature selection (we select the 100 best attributes) with only BRCA & KIRC samples in the dataset (classes with the highest number of instances). In the replication phase, we developed binary classifiers with BRCA samples and all samples from one of the other cancer types out of BRCA and KIRC. Therefore, we developed a classifier with the selected attributes in the discovery phase with the BRCA and COAD samples, another classifier with the BRCA and LUAD samples, and another one with the BRCA and PRAD samples. Next, to assess the effectiveness of the feature selection method in the replication phase, we assumed that the method would effectively replicate the genes selected in the discovery phase if, when these attributes are used to distinguish between BRCA and other cancers (instead of KYRC), we did not observe statistically significant differences in the performance of these classification models for the classification model obtained in the discovery phase. We use the statistical test phase for that purpose. For each replication cancer type (i.e., COAD, LUAD and PRAD), we use separately a paired t-test on the null hypothesis that assumes that there is no difference between means of estimates of the percentage of correct samples classified

TABLE 12. Average performance evaluation with gene expression cancer RNA-Seq dataset compared to attribute subset evaluation feature selection methods with multi-objective evolutionary search strategy, 10-fold cross-validation, 10 repetitions.

Method	*Attributes*	*NaiveBayes*	*MLP*	*SVM*	*kNN*	*RIPPER*	*C4.5*	*Random Forest*	*ZeroR*
MOEA-CFS	288	99.23	92.86	**100.00**	**100.00**	95.28	96.58	99.85	37.45
MOEA-Consistency	4	96.93	89.25	96.92	96.75	93.97	95.88	97.20	37.45
Pairwise Correlation ($q = 100$)	100	99.44	98.30	**100.00**	**100.00**	97.53	97.90	99.65	37.45
Pairwise Consistency ($q = 100$)	100	99.79	97.81	**100.00**	99.78	97.09	97.59	99.76	37.45
Pairwise Correlation ($q = 200$)	200	99.99	98.01	99.96	99.78	97.94	98.12	99.65	37.45
Pairwise Consistency ($q = 200$)	200	99.55	97.58	**100.00**	99.75	97.23	97.23	99.76	37.45

TABLE 13. Wins − losses ranking tests including attribute subset evaluation feature selection methods with multi-objective evolutionary search strategy, 10-fold cross-validation, 10 repetitions.

Method	*wins*	*losses*	*dif.*	*Rank*
MOEA-CFS	4	8	-4	5
MOEA-Consistency	0	30	-30	6
Pairwise Correlation ($q = 100$)	**9**	**0**	**9**	**2**
Pairwise Consistency ($q = 100$)	7	0	7	4
Pairwise Correlation ($q = 200$)	**10**	**0**	**10**	**1**
Pairwise Consistency ($q = 200$)	8	0	8	3

TABLE 14. Average performance evaluation with gene expression cancer RNA-Seq dataset compared to embedded feature selection methods and the top three statistically best classifiers with all attributes, 10-fold cross-validation.

Classifier	*Per cent correct*	*Attributes*
SVM – Pairwise Correlation	**100.00**	100
SVM – Pairwise Consistency	**100.00**	100
kNN – Pairwise Correlation	**100.00**	100
SVM – Pairwise Consistency	**100.00**	200
SVM	99.88	20,531
kNN	99.88	20,531
Random Forest	99.38	20,531
C4.5 (embedded)	98.00	20,531 → 5
LASSO (embedded)	99.50	20,531 → 365

TABLE 15. Mean runtimes (minutes).

Method	*Runtime*	*Rank*
Clustering Variation	0.83	5
Chi Squared	0.07	3
Correlation	0.01	1
Information Gain	0.06	2
Relief	1.34	6
Significance	0.07	4
MOEA-CFS	1216.41	12
MOEA-Consistency	791.10	11
Pairwise Correlation	414.25	10
Pairwise Consistency	384.54	9
C4.5	6.27	7
LASSO	45.45	8

between classifiers built to distinguish between BRCA and KIRC and between BRCA and the replication cancer type, when using the genes obtained at the discovery phase. And for each classifier experiment, we use 10-fold cross-validation with 10 repetitions. We have used SVM, kNN and Random Forest, the top 3 performing classifiers obtained earlier, and the accuracy results are shown in Tables 16 and 17. No statistical test performed on phase three yielded a significant p-value, therefore we could not reject the null hypothesis. Thus, we conclude that the method adequately replicates the genes with discriminative power to identify BRCA genes.

TABLE 16. External validation results of ***Pairwise Correlation*** method.

Classifier	*BRCA-KIRC*	*BRCA-COAD*	*BRCA-LUAD*	*BRCA-PRAD*
SVM	99.98	99.98	99.80	100.00
kNN	99.96	99.98	99.64	100.00
Random Forest	99.87	99.43	99.71	100.00

TABLE 17. External validation results of ***Pairwise Consistency*** method.

Classifier	*BRCA-KIRC*	*BRCA-COAD*	*BRCA-LUAD*	*BRCA-PRAD*
SVM	99.78	100.00	99.71	100.00
kNN	99.78	99.95	99.07	100.00
Random Forest	99.93	99.66	99.36	100.00

6) VISUALIZATION OF RESULTS

To verify that class separability is not affected by feature selection, in this section, we show the t-SNE visualization before and after making FS with the *Pairwise Correlation* and *Pairwise Consistency* methods, including external validation. Fig. 3 shows t-SNE visualization of gene expression cancer RNA-Seq dataset (before and after FS) and Fig. 4 to 7 shows t-SNE visualization of the datasets corresponding to the discovery and replication phases of the external validation.

B. RESULTS WITH OTHER DATASETS

To strengthen the conclusions, in this section we show results with two other gene expression datasets, in this case for *genotype-tissue expression* (GTEx) classification. GTEx [62] is an international consortium devoted to sequencing multiple parts of the human body, including 13 different brain areas, the main organs, e.g., lung, liver, heart, skin. Humans are all control subjects from a variety of ages and sex. To this date, the GTEx transcriptomic resource is the biggest repository of human tissue RNA and DNA sequencing. We downloaded TPM (Transcript per million) values from the global expression matrix of GTEx RNA expression V7 and biological covariates of interest like sex, age, and sample tissue. We filtered out all non-brain tissue samples and kept only those genes expressed with a minimum of 0.1 TMP values over 80% of the samples, within each tissue. Then separately for each brain tissue, we regressed out of the expression, RIN, age, and sex to avoid both technical and biological biases for this specific experiment. The resultant expression values

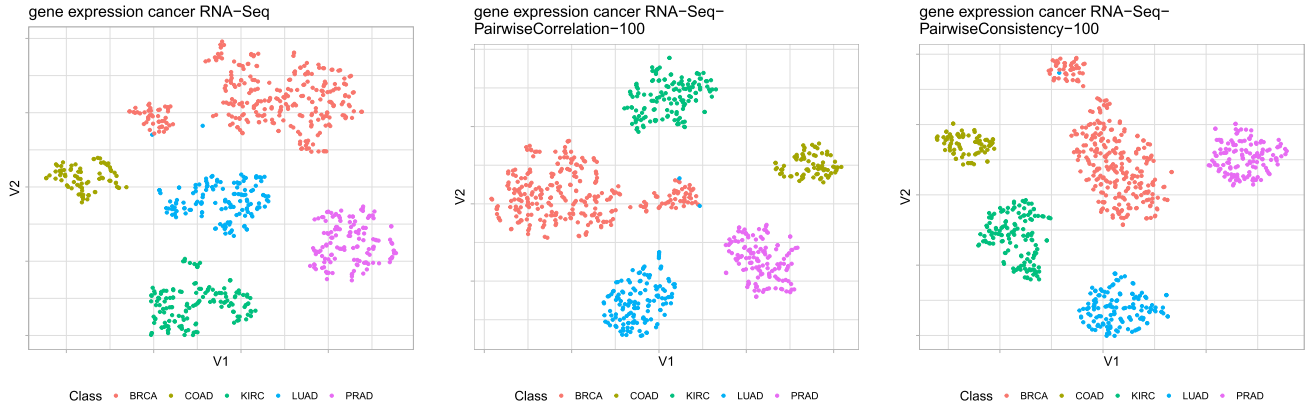

FIGURE 3. t-SNE visualization of gene expression cancer RNA-Seq dataset and reduced datasets (100 attributes) obtained with *Pairwise Correlation* and *Pairwise Consistency* methods.

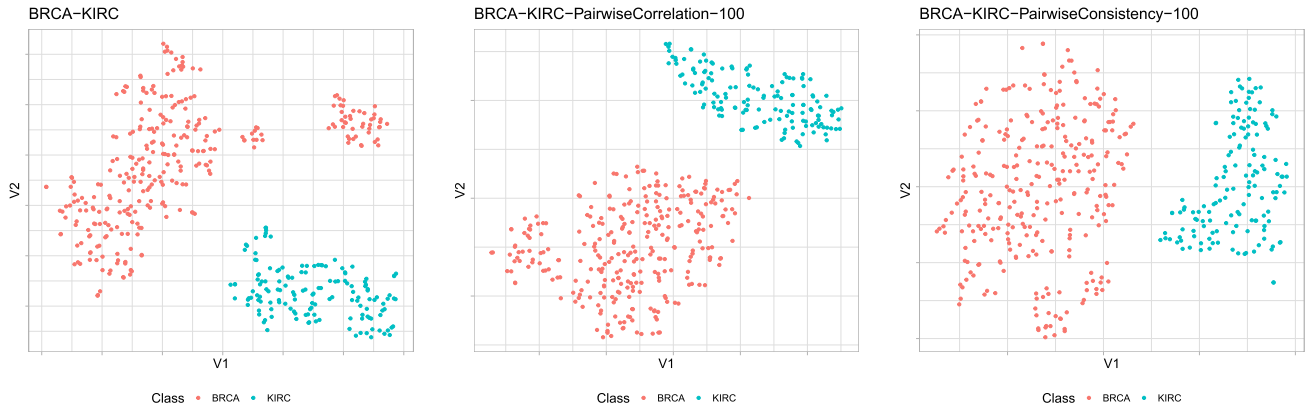

FIGURE 4. t-SNE visualization of BRCA-KIRC dataset and reduced datasets (100 attributes) obtained with *Pairwise Correlation* and *Pairwise Consistency* methods.

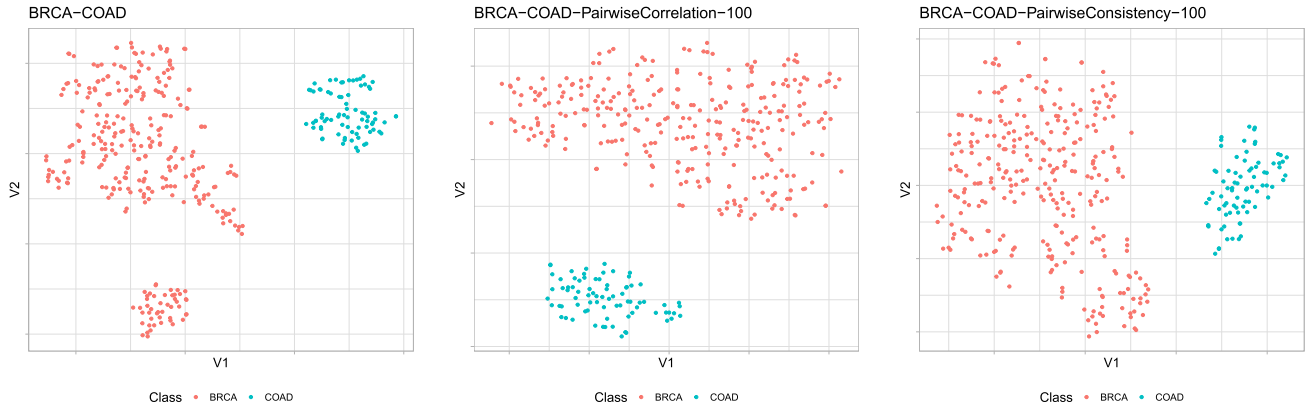

FIGURE 5. t-SNE visualization of BRCA-COAD dataset and reduced datasets (100 attributes) obtained with *Pairwise Correlation* and *Pairwise Consistency* methods.

are available in the form of an R package downloadable from GitHub.[5] The resulting dataset consists of 17863 input attributes and 1529 instances. We created two classification problems out of this dataset. The classes for the *brain tissue GTEx RNA expression* classification problem, shown in Table 18 along with the number of instances of each class, were the tissue corresponding to the specific individual sample. The classes for the *brain age GTEx RNA expression* classification problem, shown in Table 19, corresponds to 6 ranges of age for the individual.

We have performed wins-losses paired t-tests with the GTEx RNA expression datasets using the same FR and subset evaluation methods, number of attributes, classifiers, evaluation metric and validation mode that were used with the gene

[5]https://github.com/juanbot/CoExpGTExV7

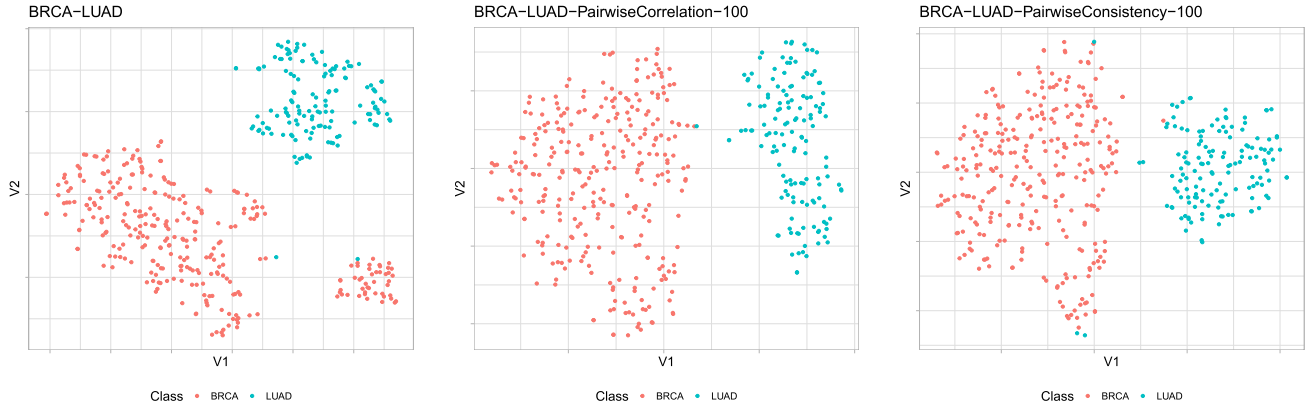

FIGURE 6. t-SNE visualization of BRCA-LUAD dataset and reduced datasets (100 attributes) obtained with *Pairwise Correlation* and *Pairwise Consistency* methods.

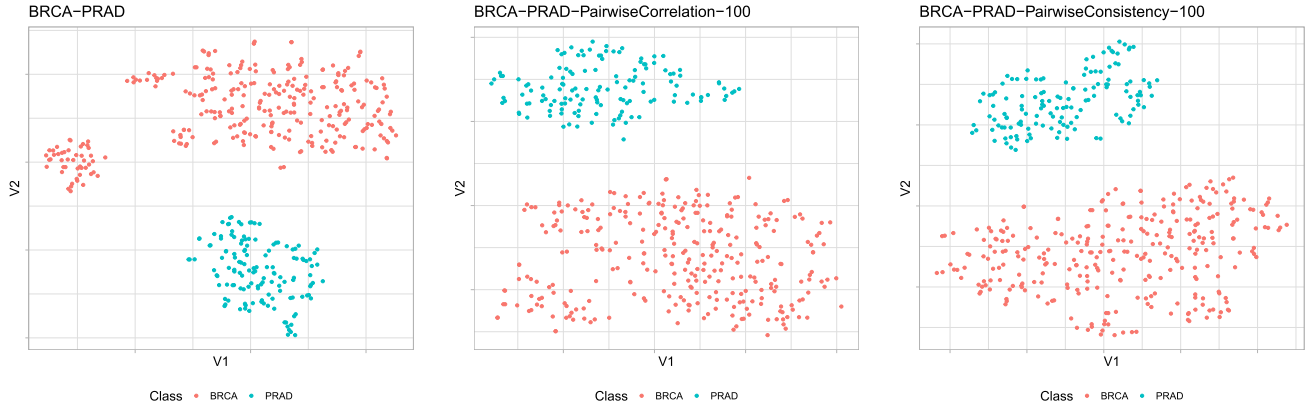

FIGURE 7. t-SNE visualization of BRCA-PRAD dataset and reduced datasets (100 attributes) obtained with *Pairwise Correlation* and *Pairwise Consistency* methods.

TABLE 18. Classes and their number of instances of the brain tissue GTEx RNA expression classification problem.

Brain tissue	*Number of instances*
BrainAmygdala	97
BrainAnteriorCingulateCortex	121
BrainCaudateBG	157
BrainCerebellarHemisphere	134
BrainCerebellum	173
BrainCortex	158
BrainHippocampus	123
BrainHypothalamus	120
BrainNucleusAccumbensBG	146
BrainPutamenBG	123
BrainSpinalCordC1	90
BrainSubstantiaNigra	87

TABLE 19. Classes and their number of instances of the brain age GTEx RNA expression classification problem.

Brain age	*Number of instances*
20-29	59
30-39	35
40-49	165
50-59	478
60-69	705
70-79	87

expression RNA-Seq dataset. Tables 20, 21 and 22 show the results of the statistical tests with the FR and subset evaluation methods, in which the best results have been marked in bold.

C. SUMMARY OF RESULTS

Finally, Tables 23 and 24 summarize the number of times each method has won minus the number of times each method has lost, taking into account the three data sets analyzed. The two best methods and their ranking positions are marked in bold in the tables. These summary tables will serve as the basis for analysing the results in the next section.

V. ANALYSIS OF RESULTS AND DISCUSSION

The following statements can be derived from the results obtained:

- The results indicate that the multivariate FR methods proposed in this paper statistically outperform the rest of the compared FR methods, both univariate and multivariate.
- If we compare the *Correlation* univariate FR method with the *Pairwise Correlation* multivariate FR method

TABLE 20. Wins − losses ranking tests for brain tissue GTEx RNA expression classification problem, 10-fold cross-validation, 10 repetitions.

	$q = 3$			$q = \log_2(n) = 14$			$q = 50$			$q = 100$		
Method	***wins***	***losses***	***dif.***	***wins***	***losses***	***dif.***	***wins***	***losses***	***dif.***	***wins***	***losses***	***dif.***
Clustering Variation	2	26	-24	10	31	-21	9	30	-21	10	30	-20
Chi Squared	20	8	12	19	11	8	22	0	22	22	1	21
Correlation	3	26	-23	0	41	-41	1	35	-34	1	35	-34
Information Gain	28	0	**28**	29	4	25	23	0	23	21	1	20
ReliefF	0	28	-28	4	34	-30	3	33	-30	4	32	-28
Significance	15	20	-5	21	6	15	15	19	-4	14	16	2
Pairwise Correlation	28	0	**28**	36	0	**36**	21	2	19	20	4	16
Pairwise Consistency	20	8	12	19	11	8	25	0	**25**	27	0	**27**

TABLE 21. Wins − losses ranking tests for brain age GTEx RNA expression classification problem, 10-fold cross-validation, 10 repetitions.

	$q = 3$			$q = \log_2(n) = 14$			$q = 50$			$q = 100$		
Method	***wins***	***losses***	***dif.***	***wins***	***losses***	***dif.***	***wins***	***losses***	***dif.***	***wins***	***losses***	***dif.***
Clustering Variation	0	10	-10	0	20	-20	1	19	-18	0	9	-9
Chi Squared	3	3	0	12	0	**12**	11	0	**11**	7	0	7
Correlation	0	15	-15	0	20	-20	0	21	-21	0	13	-13
Information Gain	3	3	0	12	0	**12**	11	0	**11**	6	0	6
ReliefF	1	5	-4	0	20	-20	0	15	-14	0	11	-11
Significance	9	0	9	12	0	**12**	11	0	**11**	8	0	**8**
Pairwise Correlation	3	4	-1	12	0	**12**	11	0	**11**	7	0	7
Pairwise Consistency	21	0	**21**	12	0	**12**	9	0	9	5	0	5

TABLE 22. Wins − losses ranking tests for GTEx RNA expression classification problems including attribute subset evaluation feature selection methods, 10-fold cross-validation, 10 repetitions.

	Brain tissue GTEx RNA			***Brain age GTEx RNA***		
Method	***wins***	***losses***	***dif.***	***wins***	***losses***	***dif.***
MOEA-CFS	8	3	5	0	4	-4
MOEA-Consistency	2	29	-27	1	0	1
Pairwise Correlation ($q = 100$)	7	6	1	3	0	**3**
Pairwise Consistency ($q = 100$)	16	0	**16**	3	0	**3**
Pairwise Correlation ($q = 200$)	6	9	-3	1	2	-1
Pairwise Consistency ($q = 200$)	10	2	8	0	2	-2

proposed in this paper, the results are clearly favourable to *Pairwise Correlation*, although both FR methods are based on the correlation metric. Basically, the difference is that the *Pairwise Correlation* method takes into account the correlation with the rest of the attributes in the evaluation of each attribute, and therefore detects its redundancy, while the *Correlation* method cannot evaluate the redundancy as it is a univariate method.

- The *ReliefF* method, even being a multivariate method, has not shown good performance in the tested gene expression classification problem, even lower than other univariate methods such as *Chi Squared*, *Information Gain* or *Significance*.
- The proposed FR methods statistically outperforms multivariate FS methods of attribute subset evaluation based on correlation and consistency, with powerful search strategies, such as multi-objective evolutionary algorithms, in shorter runtimes. This is due to the huge search space that occurs with the gene expression dataset considered in this paper. For the *cancer RNA-Seq* dataset, there are $2^{20531} = 2.8\text{e+}6180$ candidate subsets of attributes. This makes search strategies require a lot of runtime to obtain satisfactory solutions in these types of problems, which can be prohibitive in some cases. The *Pairwise Correlation* and *Pairwise Consistency* methods have evaluated a total of $n^2 = 20,531^2 = 4.2\text{e+}8$ subsets of 2 attributes, while the *MOEA-CFS* method has evaluated 3.0e+7 subsets of attributes (variable size between 1 and n) and the *MOEA-Consistency* method has evaluated 1.3e+8 subsets of attributes (also variable in size between 1 and n).
- The embedded feature selection methods *C4.5* and *LASSO* have shown poorer performance than the *Pairwise Correlation* and *Pairwise Consistency* methods proposed in this paper. The *C4.5* method selects only 5 attributes, but with a correct percentage far from 100%. The *LASSO* method selects 365 attributes without reaching 100% per cent correct. The *Pairwise Correlation* and *Pairwise Consistency* methods achieve 100% per cent correct by selecting 100 and 200 attributes respectively.
- The *Pairwise Correlation* and *Pairwise Consistency* methods proposed in this work spend more runtime than the rest of the FR methods compared, since the runtimes of the former are quadratic for to the number of attributes n, while the runtimes of the latter are linear for n.
- Both feature selection methods prove to be useful to select the best genes to identify BRCA specific cancer samples when the classifiers are interrogated with samples from unseen diseases and different patient cohorts.
- Figures 3 to 7 show that the features selected by the proposed FR methods allow classifiers to be built on a set of instances maintaining the original separability of the classes.

Below we point out the reasons why the proposed methods outperform the rest of the methods considered in the study.

TABLE 23. Ranking of the comparison with attribute evaluation feature selection methods.

	Cancer RNA-Seq		*Brain tissue GTEx RNA*		*Brain age GTEx RNA*		*Total*	
Method	***Difference***	***Rank***	***Difference***	***Rank***	***Difference***	***Rank***	***Difference***	***Rank***
Clustering Variation	-147	8	-86	6	-57	6	-250	8
Chi Squared	41	4	63	4	30	3	134	4
Correlation	-37	7	-132	8	-69	7	-238	7
Information Gain	45	3	96	2	29	4	170	3
ReliefF	6	6	-116	7	-49	5	-159	6
Significance	35	5	8	5	40	2	83	5
Pairwise Correlation	54	1	99	1	29	4	**182**	**1**
Pairwise Consistency	52	2	72	3	47	1	**171**	**2**

TABLE 24. Ranking of the comparison with attribute subset evaluation feature selection methods.

	Cancer RNA-Seq		*Brain tissue GTEx RNA*		*Brain age GTEx RNA*		*Total*	
Method	***Difference***	***Rank***	***Difference***	***Rank***	***Difference***	***Rank***	***Difference***	***Rank***
MOEA-CFS	-4	5	5	3	-4	5	-3	5
MOEA-Consistency	-30	6	-27	6	1	2	-56	6
Pairwise Correlation ($q = 100$)	9	2	1	4	3	1	13	3
Pairwise Consistency ($q = 100$)	7	4	16	1	3	1	**26**	**1**
Pairwise Correlation ($q = 200$)	10	1	-3	5	-1	3	6	4
Pairwise Consistency ($q = 200$)	8	3	8	2	-2	4	**14**	**2**

TABLE 25. Names of the classes and hyperparameters in the WEKA platform of the methods used in this paper.

Attribute evaluation methods: Evaluators	
Method name	***Hyperparameters***
weka.attributeSelection.CVAttributeEval	–
weka.attributeSelection.ChiSquaredAttributeEval	–
weka.attributeSelection.CorrelationAttributeEval	–
weka.attributeSelection.InfoGainAttributeEval	–
weka.attributeSelection.ReliefFAttributeEval	-M -1 -D 1 -K 10
weka.attributeSelection.SignificanceAttributeEval	–
weka.attributeSelection.PairwiseCorrelationAttributeEval	–
weka.attributeSelection.PairwiseConsistencyAttributeEval	–
Attribute evaluation methods: Search	
Method name	***Hyperparameters***
weka.attributeSelection.Ranker	-T -1.7976931348623157E308 -N 3
weka.attributeSelection.Ranker	-T -1.7976931348623157E308 -N 14
weka.attributeSelection.Ranker	-T -1.7976931348623157E308 -N 50
weka.attributeSelection.Ranker	-T -1.7976931348623157E308 -N 100
weka.attributeSelection.Ranker	-T -1.7976931348623157E308 -N 200
Attribute subset evaluation methods: Evaluators	
Method name	***Hyperparameters***
weka.attributeSelection.CfsSubsetEval	-P 1 -E 1
weka.attributeSelection.ConsistencySubsetEval	–
Attribute subset evaluation methods: Search	
Method name	***Hyperparameters***
weka.attributeSelection.MultiObjectiveEvolutionarySearch	-generations 300000 -population-size 100 -seed 1 -algorithm 1 -report-frequency 300000
weka.attributeSelection.MultiObjectiveEvolutionarySearch	-generations 1300000 -population-size 100 -seed 1 -algorithm 1 -report-frequency 1300000
Embedded feature selection	
Method name	***Hyperparameters***
weka.classifiers.trees.J48	-C 0.25 -M 2
weka.classifiers.mlr.MLRClassifier	-learner classif.glmnet -batch 100 -S 1
Classifiers	
Method name	***Hyperparameters***
weka.classifiers.bayes.NaiveBayes	–
weka.classifiers.functions.MLPClassifier	-N 2 -R 0.01 -O 1.0E-6 -P 1 -E 1 -S 1 -L weka.classifiers.functions.loss.SquaredError -A weka.classifiers.functions.activation.ApproximateSigmoid
weka.classifiers.functions.SMO	-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"
weka.classifiers.lazy.IBk	-K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A ẅeka.core.EuclideanDistance -R first-last"
weka.classifiers.rules.JRip	-F 3 -N 2.0 -O 2 -S 1
weka.classifiers.trees.J48	-C 0.25 -M 2
weka.classifiers.trees.RandomForest	-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
weka.classifiers.rules.ZeroR	–

1) Univariate FR methods can only detect the relevance of the attributes, but not their redundancy. The *ReliefF* method, although it is multivariate, does not detect attribute redundancy either [63]. One repeatedly noted drawback of Relief-based algorithms [64] is that they do not remove feature redundancies, i.e. they seek to

TABLE 26. Description.props of PairwiseCorrelationAttributeEval.

 # Package name
PackageName = PairwiseCorrelationAttributeEval

 # Version
Version = 1.0.1

 # Date
Date = 2021-12-15

 # Title
Title = Attribute evaluator that evaluates the worth of an attribute i by adding the merits (using CfsSubsetEval) of the attribute subsets composed of attribute i and each of the other attributes.

 # Category
Category = Attribute selection

 # Author
Author = S. Navarro, F. Jimenez, G. Sanchez, J. Palma

 # Maintainer
Maintainer = S. Navarro

 # License
License = GPL 2.0

 # Description
Description = Attribute evaluator that evaluates the worth of an attribute i by adding the merits (using CfsSubsetEval) of the attribute subsets composed of attribute i and each of the other attributes. Attributes with low correlation to other attributes and highly correlated with the class are preferred.

 # Package URL for obtaining the package archive
PackageURL = https://sourceforge.net/projects/pairwisefeatureranking/files/PairwiseCorrelationAttributeEval/PairwiseCorrelationAttributeEval1.0.1.zip/download

 # Dependencies
Depends = weka (>=3.8.3)

TABLE 27. Description.props of PairwiseConsistencyAttributeEval.

 # Package name
PackageName = PairwiseConsistencyAttributeEval

 # Version
Version = 1.0.1

 # Date
Date = 2021-12-15

 # Title
Title = Attribute evaluator that evaluates the worth of an attribute i by adding the consistency rates of the attribute subsets composed of attribute i and each of the other attributes.

 # Category
Category = Attribute selection

 # Author
Author = S. Navarro, F. Jimenez, G. Sanchez, J. Palma

 # Maintainer
Maintainer = S. Navarro

 # License
License = GPL 2.0

 # Description
Description = Attribute evaluator that evaluates the worth of an attribute i by adding the consistency rates of the attribute subsets composed of attribute i and each of the other attributes.

 # Package URL for obtaining the package archive
PackageURL = https://sourceforge.net/projects/pairwisefeatureranking/files/PairwiseConsistencyAttributeEval/PairwiseConsistencyAttributeEval1.0.1.zip/download

 # Dependencies
Depends = weka (>=3.8.3), consistencySubsetEval (>=1.0.0)

select all features relevant to the endpoint regardless of whether some features are strongly correlated with others. The proposed *Pairwise Correlation* and *Pairwise Consistency* methods detect both relevance and redundancy of the attributes.

2) The attribute subset evaluation FS methods based on correlation and consistency are multivariate methods that allow detecting relevance and redundancy of the attributes, but require a strategy to search in a space of 2^n possible subsets of attributes, which makes them inefficient for high-dimensional data since they require prohibitive runtimes to cover an acceptable percentage of search space. The proposed *Pairwise Correlation* and *Pairwise Consistency* methods require an $O(n^2 \cdot w)$ runtime to evaluate the n attributes in a dataset of w attributes and make the selection.

As a drawback of the proposed methods, we can highlight that they are more expensive in terms of runtime than the other analyzed FR methods, which is obvious since their algorithmic complexity is quadratic instead of linear. However, this is not a big drawback because FS is usually an off-line process and current advances in high-performance computing can greatly alleviate this disadvantage. What is really important is the accuracy of the classification obtained with the selected attributes. In this sense, the proposed methods are more suitable as model-agnostic methods than the rest of the compared methods, with favourable statistically significant differences in a wide range of classifiers of a diverse nature. The statistical tests carried out, with which all the methods are compared with each of the others using each of the classifiers, show that the proposed methods are in the first positions of the win-loss ranking, which means that are statistically better than the rest, for the analyzed datasets, in a larger number of classifiers.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have presented two new feature ranking methods that are especially appropriate for high-dimensional datasets, such as gene expression datasets. These methods, which we have called *Pairwise Correlation* and *Pairwise Consistency*, are filter methods based on correlation and consistency respectively that evaluate each attribute by computing the sum of the merits of all pairs of attributes formed by the attribute and each of the others, therefore they are

multivariate methods. Like any feature ranking method, *Pairwise Correlation* and *Pairwise Consistency* can also be used for feature selection. We consider that the novelty and interest of this paper lies in the fact that we redefine approaches that, while applied to single features when considering them individually as potential candidates for the final outcome, acquire a multivariate character, allowing us to identify both the relevance and the redundancy of the attributes.

We have compared the *Pairwise Correlation* and *Pairwise Consistency* methods with six feature ranking methods well known in the literature, and also with two feature selection methods of attribute subset evaluation based on correlation and consistency with the multi-objective evolutionary search strategy. Gene expression *cancer RNA-Seq* and *GTEx* RNA expression datasets have been used in the experiments. For comparisons, we used eight classification algorithms of different nature and we evaluated them with 10 repetitions of 10-fold cross-validation with the metric of per cent correct. Statistical tests have been performed to find statistically significant differences between the methods, as well as ranking wins-losses of the methods. The results of these tests place *Pairwise Correlation* and *Pairwise Consistency* in the first two positions in the ranking, outperforming univariate and multivariate feature ranking methods, attribute subset evaluation feature selection methods, and embedded methods. The *Pairwise Correlation* and *Pairwise Consistency* methods have been officially published on the WEKA platform. Since both the dataset and the software used in this work are publicly accessible, all the results shown in this paper are 100% repeatable and replicable.

We are currently analysing the best genes selected with the *Pairwise Correlation* and *Pairwise Consistency* methods for biological interpretation. As replication is crucial in medical applications, this paves the way for potentially useful uses of the algorithms in the biomarker discovery field.

APPENDIX A
NAMES OF THE METHODS AND HYPERPARAMETERS ON THE WEKA PLATFORM

All the methods used in this work are implemented in the WEKA platform. Table 25 shows the names of the classes and hyperparameters on the WEKA platform for the methods used in this article. Tables 26 and 27 show the Description.props files of the new *Pairwise Correlation* and *Pairwise Consistency* methods proposed in this document respectively.

APPENDIX B
ABBREVIATIONS

AE:	Autoencoder
ANOVA:	Analysis of Variance
BRCA:	Breast Carcinoma
CFS:	Correlation-based Feature Selection
CIFE:	Conditional Informative Feature Extraction
CMIM:	Conditional Mutual Information Maximisation
COAD:	Colon Adenocarcinoma
CSO:	Competitive Swarm Optimizer
DISR:	Double Input Symmetrical Relevance
FDR:	False Discovery Rate
FR:	Feature Ranking
FS:	Feature Selection
GE:	Gene Expression
ICAP:	Interaction Capping
JBMI:	Joint Bias Mutual Information
KIRC:	Kidney Renal Clear-cell Carcinoma
k-NN	*k*-Nearest Neighbors
LUAD:	Lung Adenocarcinoma
MAP-Elites:	Multi-dimensional Archive of Phenotypic Elites
mDSM:	modified Discretization and Selection of features based on Mutual information
ML:	Machine Learning
MI:	Mutual Information
MGS:	Mutual information-based Gene Selection
MOEA:	Multi-Objective Evolutionary Algorithm
NSGA-II:	Non-dominated Sorting Genetic Algorithm II
PANCAN:	Pan-Cancer project
PCA:	Principal Component Analysis
PRAD:	Prostate Adenocarcinoma
PSO:	Particle Swarm Optimization
RIPPER:	Repeated Incremental Pruning to Produce Error Reduction
RNA-Seq:	Ribonucleic Acid Sequencing
WER:	Weighted Ensemble of Ranks
WEKA:	Waikato Environment for Knowledge Analysis

REFERENCES

[1] V. Bolǒn-Canedo, N. Sanchez-Maroño, and A. Alonso-Betanzos, *Feature Selection for High-Dimensional Data*, 1st ed. Cham, Switzerland: Springer, 2015.

[2] P. R. Anukrishna and V. Paul, "A review on feature selection for high dimensional data," in *Proc. Int. Conf. Inventive Syst. Control (ICISC)*, Jan. 2017, pp. 1–4.

[3] V. Bolón-Canedo, A. Alonso-Betanzos, L. Morán-Fernández, and B. Cancela, *Feature Selection: From the Past to the Future*. Cham, Switzerland: Springer, 2022, pp. 11–34.

[4] A. Dabba, A. Tari, S. Meftali, and R. Mokhtari, "Gene selection and classification of microarray data method based on mutual information and moth flame algorithm," *Expert Syst. Appl.*, vol. 166, Mar. 2021, Art. no. 114012.

[5] G. Zhang, J. Hou, J. Wang, C. Yan, and J. Luo, "Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm," *Interdiscipl. Sci., Comput. Life Sci.*, vol. 12, no. 3, pp. 288–301, Sep. 2020.

[6] S. K. Baliarsingh, K. Muhammad, and S. Bakshi, "SARA: A memetic algorithm for high-dimensional biomedical data," *Appl. Soft Comput.*, vol. 101, Mar. 2021, Art. no. 107009.

[7] K. R. Kavitha, A. Prakasan, and P. J. Dhrishya, "Score-based feature selection of gene expression data for cancer classification," in *Proc. 4th Int. Conf. Comput. Methodolog. Commun. (ICCMC)*, Mar. 2020, pp. 261–266.

[8] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, 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. Bioinformat.*, vol. 9, no. 4, pp. 1106–1119, Jul./Aug. 2012.

[9] Z. Rustam and S. A. A. Kharis, "Comparison of support vector machine recursive feature elimination and kernel function as feature selection using support vector machine for lung cancer classification," *J. Phys., Conf. Ser.*, vol. 1442, no. 1, Jan. 2020, Art. no. 012027.
[10] V. Kalaimani and R. Umagandhi, "A novel wrapper FS based on binary swallow swarm optimization with score-based criteria fusion for gene expression microarray data," *Mater. Today, Proc.*, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221478532038665X, doi: 10.1016/j.matpr.2020.11.064.
[11] S. K. Panigrahi, K. Das, D. Mishra, and B. K. Veedhi, "A survey on hybridized gene selection strategies," in *Intelligent and Cloud Computing*. Cham, Switzerland: Springer, 2021, pp. 337–344.
[12] A. K. Shukla, D. Tripathi, B. R. Reddy, and D. Chandramohan, "A study on metaheuristics approaches for gene selection in microarray data: Algorithms, applications and open challenges," *Evol. Intell.*, vol. 13, no. 3, pp. 309–329, Sep. 2020.
[13] W. Zhongxin, S. Gang, Z. Jing, and Z. Jia, "Feature selection algorithm based on mutual information and lasso for microarray data," *Open Biotechnol. J.*, vol. 10, no. 1, pp. 278–286, Oct. 2016.
[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., Ser. B, Methodolog.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.
[15] S. Hameed, O. Petinrin, A. Hashi, and F. Saeed, "Filter-wrapper combination and embedded feature selection for gene expression data," *Int. J. Adv. Soft Comput. Appl.*, vol. 10, pp. 90–105, Mar. 2018.
[16] I. Kononenko, "Estimating attributes: Analysis and extensions of relief," in *Proc. Eur. Conf. Mach. Learn.*, Berlin, Germany: Springer, 1994, pp. 171–182.
[17] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, Dec. 2002.
[18] S. Maldonado and J. López, "Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification," *Appl. Soft Comput.* vol. 67, pp. 94–105, Jun. 2018.
[19] S. Gu, R. Cheng, and Y. Jin, "Feature selection for high-dimensional classification using a competitive swarm optimizer," *Soft Comput.*, vol. 22, no. 3, pp. 811–822, Feb. 2018.
[20] B. Ghaddar and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines," *Eur. J. Oper. Res.*, vol. 265, no. 3, pp. 993–1004, Mar. 2018.
[21] S. Li, K. Zhang, Q. Chen, S. Wang, and S. Zhang, "Feature selection for high dimensional data using weighted K-nearest neighbors and genetic algorithm," *IEEE Access*, vol. 8, pp. 139512–139528, 2020.
[22] C. Huang, "Feature selection and feature stability measurement method for high-dimensional small sample data based on big data technology," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–12, Sep. 2021.
[23] V. Sambhe, S. Rajesh, E. Naredo, D. Dias, M. Kshirsagar, and C. Ryan, "Multi-objective classification and feature selection of COVID-19 proteins sequences using NSGA-II and MAP-elites," in *Proc. 13th Int. Conf. Agents Artif. Intell.*, 2021, pp. 1241–1248.
[24] A. Bommert, T. Welchowski, M. Schmid, and J. Rahnenführer, "Benchmark of filter methods for feature selection in high-dimensional gene expression survival data," *Briefings Bioinf.*, vol. 23, no. 1, Jan. 2022, Art. no. bbab354.
[25] B. Trevizan and M. Recamonde-Mendoza, "Ensemble feature selection compares to meta-analysis for breast cancer biomarker identification from microarray data," in *Computational Science and its Applications ICCSA 2021*, O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blečić, D. Taniar, B. O. Apduhan, A. M. A. C. Rocha, E. Tarantino, and C. M. Torre, Eds. Cham, Switzerland: Springer, 2021, pp. 162–178.
[26] N. Pant, S. Rakshit, S. Paul, and I. Saha, "Genome-wide analysis of multi-view data of miRNA-seq to identify miRNA biomarkers for stomach cancer," *J. Biomed. Informat.*, vol. 97, Sep. 2019, Art. no. 103254.
[27] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, no. 9, pp. 1531–1555, 2004.
[28] P. E. Meyer and G. Bontempi, "On the use of variable complementarity for feature selection in cancer classification," in *Proc. Workshops Appl. Evol. Comput.*, Cham, Switzerland: Springer, 2006, pp. 91–102.
[29] A. Jakulin, "Machine learning based on attribute interactions," Ph.D. dissertation, Dept. Fac. Comput. Inf. Sci., Univerza v Ljubljani, Ljubljana, Slovenia, 2005.
[30] D. Lin and X. Tang, "Conditional infomax learning: An integrated framework for feature extraction and fusion," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2006, pp. 68–82.
[31] S. Sharmin, M. Shoyaib, A. A. Ali, M. A. H. Khan, and O. Chae, "Simultaneous feature selection and discretization based on mutual information," *Pattern Recognit.*, vol. 91, pp. 162–174, Jul. 2019.
[32] M. N. Haque, S. Sharmin, A. A. Ali, A. A. Sajib, and M. Shoyaib, "Use of relevancy and complementary information for discriminatory gene selection from high-dimensional gene expression data," *PLoS ONE*, vol. 16, no. 10, Oct. 2021, Art. no. e0230164.
[33] L. Macías-García, M. Martínez-Ballesteros, J. M. Luna-Romera, J. M. García-Heredia, J. García-Gutiérrez, and J. C. Riquelme-Santos, "Autoencoded DNA methylation data to predict breast cancer recurrence: Machine learning models and gene-weight significance," *Artif. Intell. Med.*, vol. 110, Nov. 2020, Art. no. 101976.
[34] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. thesis, Univ. Waikato, Hamilton, New Zealand, 1999.
[35] I. H. Witten, E. Frank, and M. A. Hall, "Introduction to Weka," in *Data Mining: Practical Machine Learning Tools and Techniques* (The Morgan Kaufmann Series in Data Management Systems), 3rd ed., I. H. Witten, E. Frank, and M. A. Hall, Eds. Boston, MA, USA: Morgan Kaufmann, 2011, pp. 403–406.
[36] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proc. 17th Int. Conf. Mach. Learn. (ICML)*, San Francisco, CA, USA: Morgan Kaufmann, 2000, pp. 359–366.
[37] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proc. IJCAI*, 1993, pp. 1022–1027.
[38] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, vol. 1. Cambridge, U.K.: Cambridge Univ. Press, 1988, p. 3.
[39] H. Liu and R. Setiono, "A probabilistic approach to feature selection-a filter solution," in *Proc. 13th Int. Conf. Int. Conf. Mach. Learn.*, vol. 96, 1996, pp. 319–327.
[40] H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features," in *Proc. AAAI*, vol. 91. Princeton, NJ, USA: Citeseer, 1991, pp. 547–552.
[41] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artif. Intell.*, vol. 151, nos. 1–2, pp. 155–176, 2003.
[42] C. G. A. R. Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature Genet.*, vol. 45, no. 10, pp. 1113–1120, Sep. 2013.
[43] S. Fong, J. Liang, R. Wong, and M. Ghanavati, "A novel feature selection by clustering coefficients of variations," in *Proc. 9th Int. Conf. Digit. Inf. Manage. (ICDIM)*, Sep. 2014, pp. 205–213.
[44] L. Galavotti, F. Sebastiani, and M. Simi, "Experiments on the use of feature selection and negative evidence in automated text categorization," in *Proc. Int. Conf. Theory Pract. Digit. Libraries*. Cham, Switzerland: Springer, 2000, pp. 59–68.
[45] D. Freedman, R. Pisani, and R. Purves, *Statistics (International Student Edition)*, 4th ed., R. P. Pisani, Ed. New York, NY, USA: WW Norton & Company, 2007.
[46] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *Int. J. Inf. Technol. Knowl. Manage.*, vol. 2, no. 2, pp. 271–277, 2010.
[47] A. Ahmad and L. Dey, "A feature selection technique for classificatory analysis," *Pattern Recognit. Lett.*, vol. 26, no. 1, pp. 43–56, 2005.
[48] G. I. Webb, "Naïve Bayes," in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb, Eds. Cham, Switzerland: Springer, 2017, pp. 895–896.
[49] R. Vang-Mata, *Multilayer Perceptrons: Theory and Applications* (Computer Science, Technology and Applications). New York, NY, USA: Nova Science, 2020.
[50] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.
[51] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Sep. 2006.
[52] W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn.*, San Mateo, CA, USA: Morgan Kaufmann, 1995, pp. 115–123.
[53] R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers, 1993.
[54] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[55] J. Brownlee. (Dec. 2020). *How to Estimate a Baseline Performance For Your Machine Learning Models in Weka*. [Online]. Available: https://machinelearningmastery.com/estimate-baseline-performance-machin%e-learning-models-weka/

[56] F. Jimenez, E. Marzano, G. Sanchez, G. Sciavicco, and N. Vitacolonna, "Attribute selection via multi-objective evolutionary computation applied to multi-skill contact center data classification," in *Proc. IEEE Symp. Ser. Comput. Intell.*, Dec. 2015, pp. 488–495.
[57] F. Jiménez, R. Jodár, G. Sánchez, M. Martín, and G. Sciavicco, "Multi-objective evolutionary computation based feature selection applied to behaviour assessment of children," in *Proc. Int. Conf. Educ. Data Mining (ICEDM)*, vol. 2, no. 6, 2016, pp. 1888–1897.
[58] F. Jiménez, G. Sánchez, J. M. García, G. Sciavicco, and L. Miralles, "Multi-objective evolutionary feature selection for online sales forecasting," *Neurocomputing*, vol. 234, pp. 75–92, Apr. 2017.
[59] F. Jiménez, H. Pérez-Sánchez, J. Palma, G. Sánchez, and C. Martínez, "A methodology for evaluating multi-objective evolutionary feature selection for classification in the context of virtual screening," *Soft Comput.*, vol. 23, no. 18, pp. 8775–8800, Sep. 2019.
[60] F. Jimenez, C. Martinez, E. Marzano, J. T. Palma, G. Sanchez, and G. Sciavicco, "Multiobjective evolutionary feature selection for fuzzy classification," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 5, pp. 1085–1099, May 2019.
[61] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Jan. 2002.
[62] K. G. Ardlie, "The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans," *Science*, vol. 348, no. 6235, pp. 648–660, 2015.
[63] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine Learning Proceedings 1992*, D. Sleeman and P. Edwards, Eds. San Francisco, CA, USA: Morgan Kaufmann, 1992, pp. 249–256.
[64] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *J. Biomed. Inform.*, vol. 85, pp. 189–203, Sep. 2018.

FERNANDO JIMÉNEZ received the M.D. degree in computer science from the University of Granada, in 1991, and the Ph.D. degree from the University of Murcia, in 1996, for his research on evolutionary computation applied to fuzzy transportation problems. He is currently a Professor with the Department of Communications and Information Engineering, University of Murcia. He has more than 100 publications in prestigious international journals, conferences, and book chapters. His research interests include evolutionary computation, multi-objective constrained optimization, soft computing, evolutionary fuzzy systems, machine learning, data mining, big data, and deep learning.

GRACIA SÁNCHEZ received the M.D. degree in computer science from the Polytechnic University of Valencia, Spain, and the Ph.D. degree in computer science from the University of Murcia, Spain. She is currently an Associate Professor with the University of Murcia. Her research interests include multi-objective constrained optimization, soft computing, evolutionary fuzzy systems, and data mining.

JOSÉ PALMA (Senior Member, IEEE) received the B.S. degree in computer science from the University of Las Palmas de Gran Canaria, Spain, in 1990, and the Ph.D. degree in computer science from the University of Murcia, Spain, in 1999. He has been an Associate Professor in computer science with the Department of Information Engineering and Communications and the Computer Science School, University of Murcia, since 2000, but he has been teaching in this department as an Aggregate Professor, since 1996. Before joining the University of Murcia, he worked for six years with the Department of Computer Science and Systems, University of Las Palmas de Gran Canaria. He has authored/coauthored various journal articles, book chapters, and congress papers. His research interests include medical informatics, specifically in intelligent data analysis, medical knowledge-based systems, clinical knowledge management, and ambient intelligent applications. The main techniques used in this research are machine learning, fuzzy-logic, knowledge-engineering methodologies, ontologies, and temporal reasoning.

LUIS MIRALLES-PECHUÁN worked as a full-time Researcher/Lecturer with University Panamericana, Mexico, for more than three years. In 2012, he decided to start his Ph.D. degree on creating new approaches within the online advertising world. During his Ph.D. degree, he got familiar with ML and he has published a good number of papers on topics related to how to apply ML to online advertising. After finishing his Ph.D. degree, he worked in postdoctoral levels I and II at CeADAR, UCD, and there, he has published in the Digital Forensic Conference and won the prize for the Best Student Paper. He is currently an Assistant Lecturer with TU Dublin. His current favorite topic is how to apply reinforcement learning to fight the COVID-19 pandemic and to plan the containing levels considering both the public health and the economy. Lastly, he has expertise in human activity recognition and in generalized zero-shot learning (GZSL) and applying machine learning to improve the accessibility of the websites.

JUAN A. BOTÍA received the Ph.D. degree in computational science and artificial intelligence, in March 2002. He has obtained a position with the Universidad de Murcia as a Reader, in April 2009. In 2013, he moved to London to enjoy a sabbatical period with Dr Juan C. Augusto at Middlesex University. In 2014, he joined the UKBEC Project at King's College London to work as a Research Associate with the Department of Molecular and Medical Genetics, School of Medicine, under the supervision of Mike Weale and Mina Ryten. In 2015, within the same project, he was honored to start working with John Hardy and Mina Ryten with the Department of Molecular Neuroscience, University College London, until mid-2017, when he returned to the Universidad de Murcia. He has been an Honorary Senior Research Fellow with the Institute of Neurology, University College London, U.K., since July 2017. He is currently a Professor in computer science and artificial intelligence with the University of Murcia, Murcia, Spain. During his period in the U.K., he started applying AI techniques to transcriptomics and genetics within the area of neurology, until now. He has been involved in research and innovation projects with a common aspect, such as applying AI and algorithm approaches to real-life problems, including domains like agriculture, multimedia content recommendation, indoor location of mobile devices, and ambient assisted living. His research interests include, from the very beginning of his career, multi-agent systems, distributed artificial intelligence, and machine learning with an emphasis on applications of AI. He is a member of the Ryten Laboratory.