IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 19, 2022, accepted June 8, 2022, date of publication June 15, 2022, date of current version June 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3183228

Exploiting Feature Selection in Human Activity Recognition: Methodological Insights and Empirical Results Using Mobile Sensor Data

MARCO MANOLO MANCA, BARBARA PES, (Member, IEEE), AND DANIELE RIBONI, (Member, IEEE)
Dipartimento di Matematica e Informatica, Università degli Studi di Cagliari, 09124 Cagliari, Italy

Corresponding author: Barbara Pes (pes@unica.it)

This work was supported by the Fondazione di Sardegna, under Project ADAM (L.R. 7 agosto 2007, n°7, annualità 2018) Grant CUP F74I19000900007 and Project ASTRID (L.R. 7 agosto 2007, n°7, annualità 2020) Grant CUP F75F21001220007.

ABSTRACT Human Activity Recognition (HAR) using mobile sensor data has gained increasing attention over the last few years, with a fast-growing number of reported applications. The central role of machine learning in this field has been discussed by a vast amount of research works, with several strategies proposed for processing raw data, extracting suitable features, and inducing predictive models capable of recognizing multiple types of daily activities. Since many HAR systems are implemented in resource-constrained mobile devices, the efficiency of the induced models is a crucial aspect to consider. This paper highlights the importance of exploiting dimensionality reduction techniques that can simplify the model and increase efficiency by identifying and retaining only the most informative and predictive features for activity recognition. More in detail, a large experimental study is presented that encompasses different feature selection algorithms as well as multiple HAR benchmarks containing mobile sensor data. Such a comparative evaluation relies on a methodological framework that is meant to assess not only the extent to which each selection method is effective in identifying the most predictive features but also the overall stability of the selection process, i.e., its robustness to changes in the input data. Although often neglected, in fact, the stability of the selected feature sets is important for a wider exploitability of the induced models. Our experimental results give an interesting insight into which selection algorithms may be most suited in the HAR domain, complementing and significantly extending the studies currently available in this field.

INDEX TERMS Feature selection methods, human activity recognition, machine learning algorithms, mobile sensor data.

I. INTRODUCTION

Sensor-based *Human Activity Recognition* (HAR) is a growing research field that deals with automatically identifying the activities a user is performing based on the analysis of data collected from a variety of sensors [1]. HAR systems have several important applications in different areas including ambient assisted living [2], physical training [3], [4], activity monitoring for health assessment [5], assistance for child and elderly care [6] as well as assistance for people with cognitive disorders [7].

In particular, sensors fitted in mobile devices (accelerometers, gyroscopes, etc.) have now reached widespread adoption and allow to envisage intelligent monitoring systems that can seamlessly track daily activities, in a non-intrusive way, and help users to make better decisions about their future actions [8]–[10]. However, such an automatic decision support capability is still limited, despite the increasing capacity of processing data from smart devices. Furthermore, the exploitation of HAR systems in real-world scenarios not only demands high recognition accuracy but also poses multiple challenges in terms of power consumption and robustness with respect to different context conditions [11], [12], soliciting further research efforts in this field.

The associate editor coordinating the review of this manuscript and approving it for publication was Dongxiao Yu.

In recent years, several machine learning approaches have been explored for the automatic classification of daily living activities based on mobile sensing [13]–[18]. The overall process involves different steps, including the cleansing of raw data to remove noise and artifacts, and the extraction of high-level features that can be useful to discriminate among the considered activities [19]. A common methodology for feature extraction relies on segmenting the sensor data, e.g., the tri-axial acceleration signals, into time windows that are subsequently mapped, through proper functions, into a set of meaningful features. Different types of mapping approaches have been explored based on the application scenario [20], resulting in time-domain, frequency-domain, or other types of features that can be finally fed to a classifier to induce the HAR model. Automatic feature extraction based on deep learning methods has also been recently investigated [11], [21]. Both the features' definition and the choice of the classifier may significantly affect the performance of the recognition system, as witnessed by a vast amount of literature in the field [5], [8], [20].

The computational efficiency of the HAR system is another important aspect to consider when dealing with mobile devices that may have limitations in terms of processing capability as well as energy consumption. From this point of view, it may be convenient to contain the dimensionality of the feature vectors used at the classification stage, as discussed in some recent studies [22]–[24]. The feature extraction process can indeed result in a large number of features, especially in a multi-sensor system where the feature sets generated from the single sensors can be fused to create feature vectors of high dimensionality [25]. In such a scenario, it can be useful to apply automatic techniques to identify and select the most important features for prediction, in order to simplify the activity recognition models, decrease overfitting, and reduce computations. Indeed, it has been observed that not all the extracted features have the same importance in the classification of physical activities [19], [26]. Some features may be redundant, weakly relevant, or even irrelevant/noisy for a specific task, suggesting that feature selection techniques could be effectively employed to reduce the data dimensionality and improve the HAR system's efficiency without compromising the final prediction accuracy.

However, not much research has so far explored the potential of feature selection in this field. Most emphasis has been given to investigating which classification methods and strategies may be best suited for inducing the HAR models [5], [20], [27], [28], while few studies have compared the impact of different feature selection algorithms on such models [29], gaining insight into which heuristics may be most effective in selecting reduced subsets of discriminative features. To make a contribution in this direction, we present here an extensive comparative study that encompasses several selection approaches and investigates their behavior on typical HAR datasets extracted from mobile devices, like smartphones and smartwatches, where recognition performance and computational efficiency need to be jointly optimized.

More in detail, extending our previous research in this field [30], this work provides a two-fold contribution. First, a general methodological framework is presented that allows evaluating the effectiveness of the feature selection process along with two directions: *i)* the capacity of the algorithm of identifying the most important features for activity prediction, and *ii)* the stability of the selected feature subsets, i.e., their robustness to perturbations in the input data. This way we can better understand the suitability of a given selection method for the considered application scenario. Indeed, identifying feature subsets that are both highly predictive and stable is important for a wider exploitability of the induced models, as highlighted in recent literature [31], [32].

Leveraging such a framework, we carried out an experimental evaluation for different levels of dimensionality reduction, i.e., for different percentages of selected features, in order to find an optimal trade-off between the number of features used for prediction and the resulting classification performance. Specifically, our study includes selection methods that are representative of different heuristics: *univariate* approaches, which assess every single feature independently of the others; *multivariat*e approaches, which capture the inter-dependencies among the features; *filter* approaches, which carry out the selection process without interacting with the learning algorithm; *embedded* approaches, which exploit the features' weights derived by a proper classifier to assess the relevance of the features. Such a comprehensive evaluation has been performed in conjunction with learning algorithms that have proven highly effective in this domain, such as *Support Vector Machines* and *Random Forest*.

The experimental study has been conducted on five HAR datasets extracted from mobile sensor data [33]–[37], both using a single sensor type (accelerometer) or different types of sensors (accelerometer, gyroscope, magnetometer). Further, the considered benchmarks present different levels of dimensionality, ranging from about 150 features to over a thousand features, which has allowed us to explore the behavior of the considered selection algorithms across different feature spaces.

Overall, the results of our experiments show that the feature selection process can reduce the original dimensionality to a great extent without any degradation in the final recognition performance, which confirms the importance of introducing an automatic dimensionality reduction step into any mobile sensing processing pipeline. By jointly evaluating the predictive performance and the selection stability, we also obtained some interesting insights into which methods may be best suited in the considered domain. To the best of our knowledge, this is the first work that evaluates several feature selection approaches in this field based on different real-world benchmarks. Moreover, in this work we investigate the stability of selection methods, which is neglected by most of the existing studies.

The remainder of this paper is structured as follows. Section II gives some background concepts on feature selection and discusses its applications in the HAR field.

Section III describes the adopted methodological framework and the specific selection algorithms chosen for the experimental study. The characteristics of the considered datasets are illustrated in Section IV. Section V presents the experimental analysis and Section VI further discusses the main findings. Finally, Section VII gives some concluding remarks as well as directions for future work.

II. BACKGROUND CONCEPTS AND RELATED WORK

In the last two decades, several research efforts have focused on devising proper methods for handling high dimensional datasets [38]. Feature selection plays an important role in such a context as it can discard irrelevant and redundant information, as well as noisy factors, with significant benefits in terms of computational efficiency, model interpretability and data understanding [39]. As summarized below, a wide variety of feature selection methods can be found in the literature, with some promising applications also in the HAR field [22]–[24], [29], [40]–[51], [53]–[55].

A. BACKGROUND ON FEATURE SELECTION

In the context of supervised learning tasks, like those considered in this paper, the available selection techniques can be broadly distinguished into three categories [39], [56]:

- *Filters* methods, that conduct the selection process as a pre-processing step, without interacting with the learning algorithm used at the model induction stage, thus leading to a classifier-independent selection outcome. This can involve the *individual evaluation* of every single feature, based on its correlation with the class attribute, or the *evaluation of subsets* of features, within which the reciprocal correlation among the features is also considered to minimize redundancy (but at an increased computational cost).
- *Wrapper* methods, that search for the feature subset that can optimize the predictive performance of a given classifier. In this case, the learning algorithm itself is employed as an evaluation function to assess each candidate subset of features. The computational cost of the selection process is hence dependent on the classifier's intrinsic efficiency, as well as on the search strategy used to build the candidate subsets (e.g., an evolutionary search or a greedy stepwise search), with an overall burden generally higher than the one of filter methods.
- *Embedded* methods, that leverage the intrinsic capacity of some classification algorithms (e.g., *Support Vector Machines* classifiers) to assign weights to the features, without requiring a systematic search through different candidate subsets, as in the case of wrappers. In terms of computational cost, this approach often provides a reasonable trade-off between filters and wrappers, with results that have proven quite satisfactory in multiple scenarios.

Several studies have investigated the potential and the drawbacks of the different selection methods proposed so far [57], [58]. Due to their reduced computational requirements, filter and embedded methods have found wider adoption in high-dimensional problems, with a variety of algorithms available that support both *univariate* and *multivariate* selection processes, as better discussed in section III. Hybrid and ensemble techniques, that properly integrate different selection methods, have also been studied with promising results in recent years [39], [59], [60].

B. FEATURE SELECTION IN THE HAR FIELD

Feature selection methods falling in the filter category have been generally preferred in the HAR field. For example, [40] and [41] have investigated the use of *MRMR* and *CFS* filters [38], which are designed to search for subsets of features that are highly correlated with the target class but not correlated with each other. Similarly, the inter-dependencies among the features are considered in [42], where a game theory-based feature selection (*GTFS*) method is proposed to optimize the size of the selected feature subsets. Such a subset-oriented evaluation tries to reduce redundancy and can be an effective and viable solution when the original dimensionality is not too high.

A more efficient ranking-based filter is exploited in [43], where every single feature is weighted according to an information theoretical criterion, known as *Information Gain*, with an overall ordering of features based on the resulting weights. Such an approach allows discarding the features that are less useful in discriminating the target class and has proven well suited even in the presence of very high dimensionalities. Ranking-based filters have also been applied in [44], [45], where the *Chi-Squared*, *Fisher score*, and *ReliefF* methods are employed to weight the features and arrange them in decreasing order of relevance. A comparison between ranking-based and subset-oriented filters is presented in [23], which emphasizes that the filter-selected, classifier-independent, feature subsets have potentially broad exploitability in smartphone-based HAR.

Fewer applications can be found in this field for the embedded methods and the wrapper methods [46]–[48]. Specifically, given the higher computational cost of wrappers, they have been mainly applied after preliminarily reducing the data dimensionality by means of a filter, as in [24] and [49]. Some direct comparisons between filter and wrapper approaches are presented in [22], [50], [51]; in such studies, however, the dimensionality of the original feature space is relatively low, which can make acceptable the higher computation time required by wrappers.

An interesting line of research recently explored in the HAR field relies on the use of nature-inspired and swarm intelligence algorithms [52] for optimizing the feature subset selection process within a wrapper model approach. In particular, [53] proposes a new method that involves the hybridization of two algorithms, namely the *gradient-based optimizer* (*GBO*) and the *grey wolf optimizer* (*GWO*). A *binary firefly* algorithm (*BFA*) is used in [54], in combination with a *GTFS* filter that preselects potentially interesting features.

The integration of *bee swarm optimization* (*BSO*) and reinforcement learning is explored in [55], with a comparison with other swarm-based methods.

Despite an increasing amount of research pointing out the benefits of feature selection in HAR applications, there is a lack of comparative studies that extensively evaluate the strengths and weaknesses of the different selection approaches in this field. At the time of writing, the largest experimental comparison is presented in [29], where ten different selection algorithms (seven filters, two wrappers and one embedded method) are evaluated on a single sensor dataset involving more than 200 attributes; interestingly, the feature subsets selected using efficient filter methods are found to outperform those produced by wrappers that, although potentially capable of yielding superior results, are more prone to the problem of overfitting.

Finally, to the best of our knowledge, only the impact of feature selection on the performance of HAR models has been considered so far, without investigating the stability of the selection process, i.e., its sensitivity to changes in the input data, which may critically affect the robustness of the induced models. Taking such an aspect into account, this work presents a wide comparative analysis that complements the studies available in this field, encompassing several selection methods and several benchmarks, as detailed in the following sections.

III. METHODOLOGICAL FRAMEWORK

Our methodological framework is meant to be general enough to be implemented with different selection methods as well as different learning algorithms. Further, as anticipated above, it involves evaluating both the predictive power and the stability of the selected feature subsets, which is important to understand the extent to which these subsets can be truly relevant for the task at hand, regardless of the specific composition of the training data. All the steps of the adopted methodology are outlined in what follows, along with a description of the specific methods and settings chosen for the comparative analysis.

A. RANKING-BASED FEATURE SELECTION AND PERFORMANCE EVALUATION

As a general framework for a wide comparison, we chose a ranking-based selection approach [39] that is flexible enough to encompass the use of *filter* methods, that assign weights to the features based on their degree of correlation with the class (i.e., the activity to be predicted), and *embedded* methods, that rely on the features' weights derived by a proper learning algorithm.

The assigned weights, regardless of how they are computed, can be used to obtain a *ranked list* in which the N features of the data at hand (D) appear in decreasing order of relevance, i.e., from the most important (rank 1) to the least important (rank N), as schematized in Figure 1. Such a list can then be cut at a suitable threshold point (n) to select a subset of highly relevant features, i.e., the n top-ranked ones.

Considering only these features, an activity recognition model can be induced from D by training any suitable classifier. The performance of the resulting model is evaluated on a separate set of test records that are structured to contain only the features previously selected from D (to avoid any selection bias, in fact, the test records must not be used in the feature selection process).

The best level of dimensionality reduction, i.e., the optimal value of the cut-off threshold in Figure 1, may vary depending on the specific task at hand. The effect of modifying such a threshold is explored experimentally in our study, which encompasses different values of n and evaluates their impact on the final recognition performance. This approach allows discarding the unnecessary features in a cost-effective way (especially when the data dimensionality makes impractical the direct adoption of *wrapper*-based search strategies).

B. STABILITY EVALUATION

For a stable selection method, we expect to obtain (almost) the same outcome when the original set of training instances is somewhat perturbed (e.g., randomly removing a given percentage of records) [61].

Evaluating stability essentially involves two aspects [62]: *(i)* a procedure to create multiple sample sets from the available data, and *(ii)* a consistency index to quantify the sensitivity of the selection process to sample variation. More in detail, given a dataset D with R instances and N features, a number K of reduced datasets D_i ($\mathrm{i} = 1, 2, \ldots, K$) are drawn, each containing a fraction f of the original instances. The chosen selection algorithm is then applied to each D_i, obtaining an output O_i ($\mathrm{i} = 1, 2, \ldots, K$) that may depend on D_i's specific composition. A proper similarity measure is finally used to assess the pairwise similarity between the outputs O_i: the more their average similarity, the more stable the selection method.

In our framework, each output O_i takes the form of a feature subset S_i containing n of the original features (selected according to the approach explained previously), for a total of K subsets of the same size. To evaluate how similar these subsets are to each other, we rely on a consistency index known as *Kuncheva measure* [63], which has proved to be suitable in the context of high dimensional problems. Specifically, given a pair of subsets S_i and S_j, their similarity is measured as follows:

$$sim_{ij} = \frac{|S_i \cap S_j| - n^2/N}{n - n^2/N} \tag{1}$$

where $|S_i \cap S_j|$ is the number of features that are common to S_i and S_j. The similarity sim_{ij} essentially measures the degree of overlap between the two subsets, with a proper correction reflecting the probability that a feature is included in both subsets simply by chance [31].

The resulting similarity values are then averaged across all pair-wise comparisons to assess the overall degree of

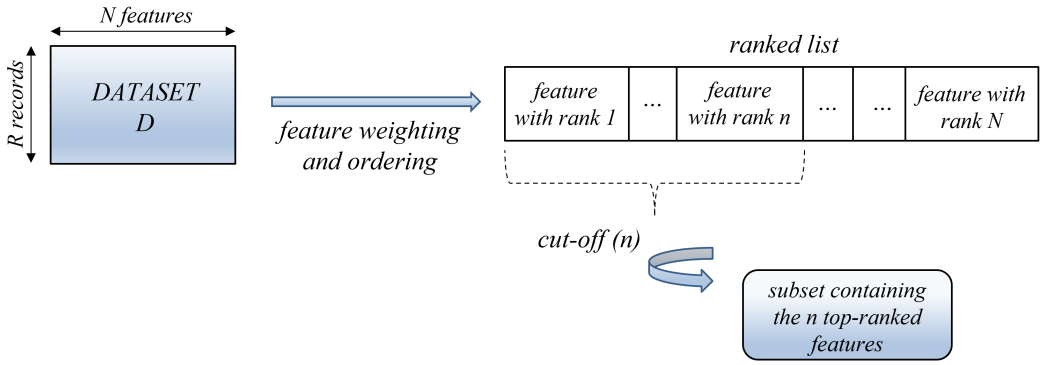

FIGURE 1. The adopted ranking-based selection approach.

consistency among the K subsets:

$$sim_{avg} = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} sim_{ij} \tag{2}$$

This average similarity can be assumed as a measure of the stability level of the selection process. Since sim_{ij} and sim_{avg} may vary in dependence on the size n of the selected subsets, our experimental study investigates the stability trend for feature subsets of increasing size, as shown in Section V.

C. METHODS AND SETTINGS

For implementing the methodological framework presented above, we considered some popular selection algorithms that are representative of different feature weighting paradigms. In particular, we employed five *univariate methods*, that weigh every single feature independently of the others, and five *multivariate* methods, that are able to capture the inter-dependencies among the features.

Specifically, among the univariate techniques, we chose:

- *Chi Squared* (χ^2), that leverages the well-known chi-squared statistic to evaluate how relevant a feature is with respect to the class [57]. Specifically, once a feature has been discretized into I intervals, its χ^2 value is obtained as:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{C} \frac{(A_{ij} - \frac{R_i \cdot B_j}{R})^2}{\frac{R_i \cdot B_j}{R}} \tag{3}$$

  where R is the total number of instances, C the number of classes, R_i the number of instances in the *ith* interval, B_j the number of instances in the *jth* class, and A_{ij} the number of instances in the *ith* interval and *jth* class.
- *Information Gain* (*IG*), that measures the extent to which the class entropy decreases when the value of a given feature is known: the greater the decrease in entropy, the more discriminative the feature [64]. Namely, by denoting as H the entropy function, we can derive the *IG* value for a feature X as:

$$IG(X) = H(Y) - H(Y|X) \tag{4}$$

  where $H(Y)$ is the entropy of the class Y before observing X, while $H(Y|X)$ is the conditional entropy of Y given X [65].
- *Symmetrical Uncertainty* (*SU*) and *Gain Ratio* (*GR*), that in turn rely on the *IG* measure but include suitable correction factors that try to compensate for the *IG*'s bias toward features with more values [66]. Specifically, after computing the *IG* value for a feature X, the corresponding *SU* and *GR* values are obtained as follows:

$$SU(X) = \frac{2 \cdot IG(X)}{H(X) + H(Y)} \tag{5}$$

$$GR(X) = \frac{IG(X)}{H(X)} \tag{6}$$

  where $H(X)$ and $H(Y)$ denote the entropy of, respectively, the feature X and the class Y.
- *OneR* (*OR*), that weights each feature based on the accuracy of a simple classification rule built on that feature, according to the approach originally proposed in [67]. More in detail, for each of the available features, the algorithm creates one rule by determining the most frequent class for each feature's value (a rule is simply a set of attribute values bound to their majority class). The prediction accuracy of each rule is then computed, and the features are ranked according to the quality of the corresponding rules.

Among the multivariate techniques, we considered two *Relief*-based selectors [68] and three *SVM*-based selectors [69]. More in detail, we chose:

- *ReliefF* (*RF*), that evaluates the strength of the features according to their ability to discriminate between data instances that are near to each other (nearest neighbors) in the feature space. Basically, in the original two-class formulation, a sample instance R_i is extracted from the training set, and its features' values are compared to the corresponding values of the instance's nearest *hit H* (neighbor from the same class) and *miss M* (neighbor from the opposite class). A weight W is then iteratively computed for each feature X, starting from an initial

value $W(X) = 0$:

$$W(X) := W(X) - \frac{diff(X, R_i, H)}{r} + \frac{diff(X, R_i, M)}{r} \tag{7}$$

where r is the number of randomly drawn instances and *diff* is a function that computes the difference between the value of X for R_i and H as well as for R_i and M. The underlying assumption is that a "good" feature should have the same value for data points of the same class and different values for data points of different classes. Such a binary formulation can be easily extended to deal with multi-class problems [68]. In turn, *ReliefF-weighted* (*RFW*) adopts a similar strategy but weighting the neighbors by their distance.

- *SVM-AW*, that relies on a linear *SVM* classifier, which has an embedded capability of assigning a weight to each feature based on how it contributes to the hyperplane decision function induced by the classifier. This function can indeed be written as follows:

$$f(\boldsymbol{X}) = \boldsymbol{W} \cdot \boldsymbol{X} + b \tag{8}$$

where $\boldsymbol{X}$ is the N-dimensional vector of input features, $\boldsymbol{W}$ is a weight vector, and b is a bias constant. The weight W_j assigned to the *jth* feature can be interpreted as a measure of the strength of the feature; specifically, the *SVM-AW* algorithm considers the absolute value of this weight (*AW*) [69].

- *SVM-RFE*, that, in turn, exploits a linear *SVM* but adopts a *recursive feature elimination* strategy that iteratively removes the features with the lowest weights and repeats the weighting process on the remaining features. In our study, two versions of this approach are evaluated: *RFE10*, where the percentage p of features removed at each iteration is set to 10%, and *RFE50*, where this percentage is 50% (in the special case where $p = 100\%$, *SVM-RFE* reduces to *SVM-AW* as no iteration occurs).

As regards the computational complexity of the above techniques, it depends on both the number of features (N) and the number of instances (R). In particular, it can be shown that the number of operations is of the order of $N \cdot R$ for the univariate approaches [70], while the multivariate approaches have a higher computational cost. Indeed, in the worst case, the number of operations is of the order of $N \cdot R^2$ for the *Relief*-based methods while it is of the order of $max(N, R) \cdot R^2$ for *SVM-RFE*. However, efficient implementations exist that optimize the nearest-neighbor calculations involved in *ReliefF* as well as the kernel-matrix calculations involved in *SVM-RFE* [71].

Furthermore, note that some of the above techniques (χ^2, *IG*, *GR*, *SU*, *RF*, and *RFW*), which only rely on the data's intrinsic characteristics, fall in the category of filter methods, while others (*OR*, *SVM-AW*, *RFE10*, and *RFE50*) leverage the features' weights derived by a suitable classifier and can be thus categorized as embedded methods. Irrespective of the specific algorithm used to derive the features' weights, the final selection is carried out according to the approach shown in Figure 1, i.e., by retaining a number n of top-ranked features; such a number is varied in our experiments encompassing different percentages of selected features, from 5% to 90%.

As learning algorithms to induce the activity recognition models, we chose, after a series of preliminary experiments, an *SVM* classifier with a polynomial kernel of degree 2 and a *Random Forest* classifier, which proved to be a suitable option for the considered benchmarks (described in section IV). For both classifiers, as well as for the different selection methods, we leveraged the implementations provided by the *WEKA* machine learning library [72].

More in detail, we trained the *SVM* classifier using the well-known *Sequential Minimal Optimization* (*SMO*) algorithm [73]. For the *Random Forest* classifier, we relied on 100 unpruned trees, each built using $\log_2(n) + 1$ random features at the splitting stage, according to commonly adopted settings [74]. The *WEKA ChiSquaredAttributeEval*, *InfoGainAttributeEval*, *SymmetricalUncertAttributeEval*, *GainRatioAttributeEval*, and *OneRAttributeEval* were used to implement the univariate methods χ^2, *IG*, *SU*, *GR*, and *OR*, respectively. For the *Relief*-based methods, we employed the *ReliefFAttributeEval* function, with and without instance weighting (for *RFW* and *RF* respectively). Finally, for the *SVM*-based selection methods, i.e., *SVM-AW* and *SVM-RFE*, we exploited the *SVMAttributeEval* function, properly setting the percentage of features to be removed at each iteration. Each of these feature weighting functions was used in conjunction with the *Ranker* search method that allows selecting a specified number of top-ranked features.

IV. DATASETS

For our comparative study, we used five datasets meant for mobile human activity recognition. Specifically, one of these datasets contains data acquired from a fitness watch and a mobile phone, three of them contain smartphone sensor data, and the last one contains body-worn sensor data. In each dataset, data instances are labeled with the corresponding activities, which are primarily sport/fitness activities and activities of daily living. The high-level characteristics of these experimental benchmarks are summarized in Table 1.

The ***COSAR*** dataset [33], [75] was collected in experiments concerning the recognition of 10 different activities: *brushing teeth*, *climbing up and down*, *riding bycicle*, *standing still*, *jogging and strolling*, *walking downstairs and upstairs*, *writing on blackboard*. These activities were carried out by 6 volunteers wearing two accelerometers: the first one located inside the left pocket and the second one on the right wrist. In addition, a GPS receiver in the left pocket tracked the person's location. Overall, the dataset consists of 5 hours of activity data, sampled at a frequency of 16Hz, and each activity instance has a time extension of 1 second, for a total of 18000 instances (divided into 13500 training records and 4500 test records, as reported in Table 1).

TABLE 1. Datasets used in the experimental study.

Dataset	Number of records	Number of features	Number of classes	Sensors
COSAR (*Everywarelab Activity Recognition Dataset*)	18000 (13500 training + 4500 test)	148	10	2 accelerometers
HAR (*Human Activity Recognition Using Smartphones Dataset*)	10299 (7352 training + 2947 test)	561	6	1 accelerometer, 1 gyroscope
HAR_ALL (*Smartphone Dataset for Human Activity Recognition in Ambient Assisted Living*)	5744 (4252 training + 1492 test)	561	6	1 accelerometer, 1 gyroscope
HAPT (*Smartphone-Based Recognition of Human Activities and Postural Transitions Dataset*)	10929 (7767 training + 3162 test)	561	12	1 accelerometer, 1 gyroscope
DSA (*Daily and Sports Activities Dataset*)	9118 (6838 training + 2280 test)	1170	19	5 accelerometers, 5 gyroscopes, 5 magnetometers

The second dataset (***HAR***) [34] was collected from 30 volunteers wearing a smartphone on the waist, as described in [76]. Inertial data were acquired from the smartphone's 3-axial accelerometer and 3-axial gyroscope at 50 Hz. Each subject carried out six activities: *walking*, *walking upstairs and downstairs*, *sitting, standing and laying*. Raw data were filtered to remove noise and sampled using a 50% overlapping sliding window of 2.56 seconds, resulting in a total of 10299 instances (partitioned into 70% training data and 30% test data).

The third dataset (***HAR_ALL***) [35] is an extension of the *HAR* dataset described above; indeed, it was obtained by carrying out similar experiments and contains the same activities, as described in [77]. 30 subjects participated in the experiments, resulting in a collection of 5744 records.

In turn, the ***HAPT*** dataset [36] is an updated version of the *HAR* dataset that contains an extended set of activities, including postural transitions: *walking*, *walking upstairs and downstairs*, *sitting*, *standing and laying*, *stand-to-sit and stand-to-lie*, *sit-to-stand and sit-to-lie*, *lie-to-sit and lie-to-stand*. This benchmark, described in detail in [14], is considered especially challenging due to the highly imbalanced activity distribution (given the lower frequency of postural transitions).

The last dataset we considered is (***DSA***) [37], previously used in various research works including [78] and [79]. It contains 19 daily and sport activities: *sitting*, *standing and lying on back and on right side*, *ascending and descending stairs*, *standing and moving around in an elevator*, *walking in a parking lot and on a treadmill (in flat and 15° inclined positions)*, *running on a treadmill*, *exercising on a stepper and on a cross trainer*, *cycling on an exercise bike in horizontal and vertical positions*, *rowing*, *jumping and playing basketball*. The activities were carried out by 8 subjects and the total duration of each activity was 5 minutes per subject. Data were acquired at a sampling rate of 25 Hz, using five body-worn orientation trackers, each containing a 3-axial accelerometer, a 3-axial gyroscope and a 3-axial magnetometer. The sensor signals were divided into 5-second segments, yielding a total of 9120 instances (480 per activity).

As we can see in Table 1, the five considered benchmarks have quite different dimensionalities, due to the different numbers of sensors involved as well as to the different set of high-level features computed for each sensor signal (after segmenting it into time windows). Specifically, for each window, feature vectors were computed in the time and frequency domains (*Fast Fourier Transform* was applied to sensor signals to transform data in the frequency domain). Several statistical measures were computed for each window: some of them, e.g., mean, standard deviation, Kurtosis, and min/max values, were extracted for all five datasets, while others have been used only in some of them. Note that we maintained the features' definitions employed in the original studies in which the datasets were published, in order to avoid introducing any bias and make the experiments fully repeatable. See Appendix I for more details on the extracted features.

V. EXPERIMENTAL ANALYSIS

Leveraging the activity recognition benchmarks described in the previous section, we conducted an extensive experimental study aimed at investigating the extent to which the original feature space can be reduced without degrading the final predictive performance. The overall analysis, in terms of both capacity of discriminating the classes and selection stability, has been carried out for different levels of dimensionality reduction, according to the methodological framework detailed in Section III. The main experimental results are summarized in Figures 2, 3 and 4, each containing ten charts that compare the behavior of the considered selection methods.

Specifically, Figure 2 and Figure 3 show a comparison in terms of *F-score*, which is a performance metric widely employed in activity recognition tasks. Defined as the

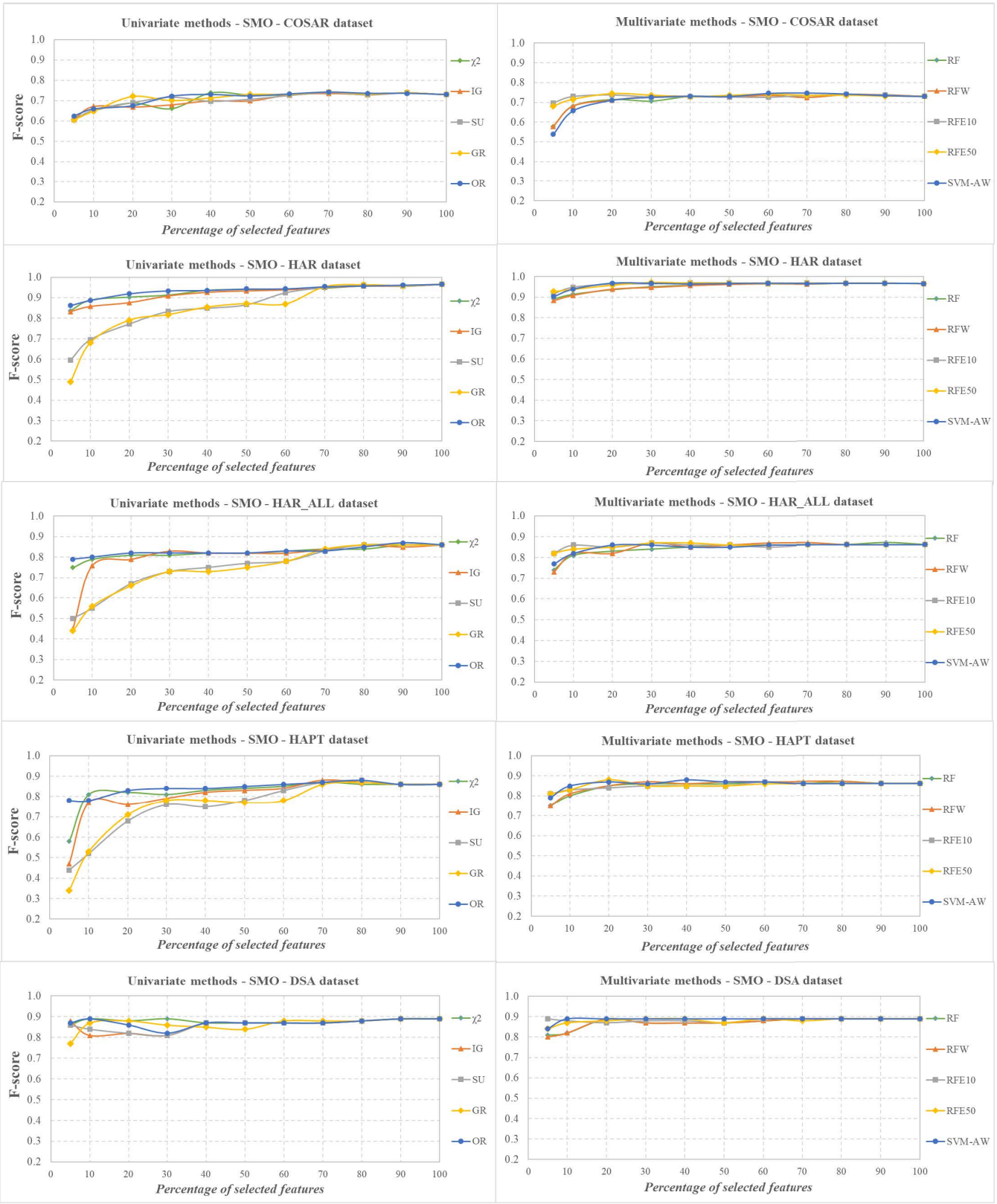

FIGURE 2. F-score performance of the *SMO* classifier, in conjunction with the univariate (on the left) and the multivariate (on the right) selection methods.

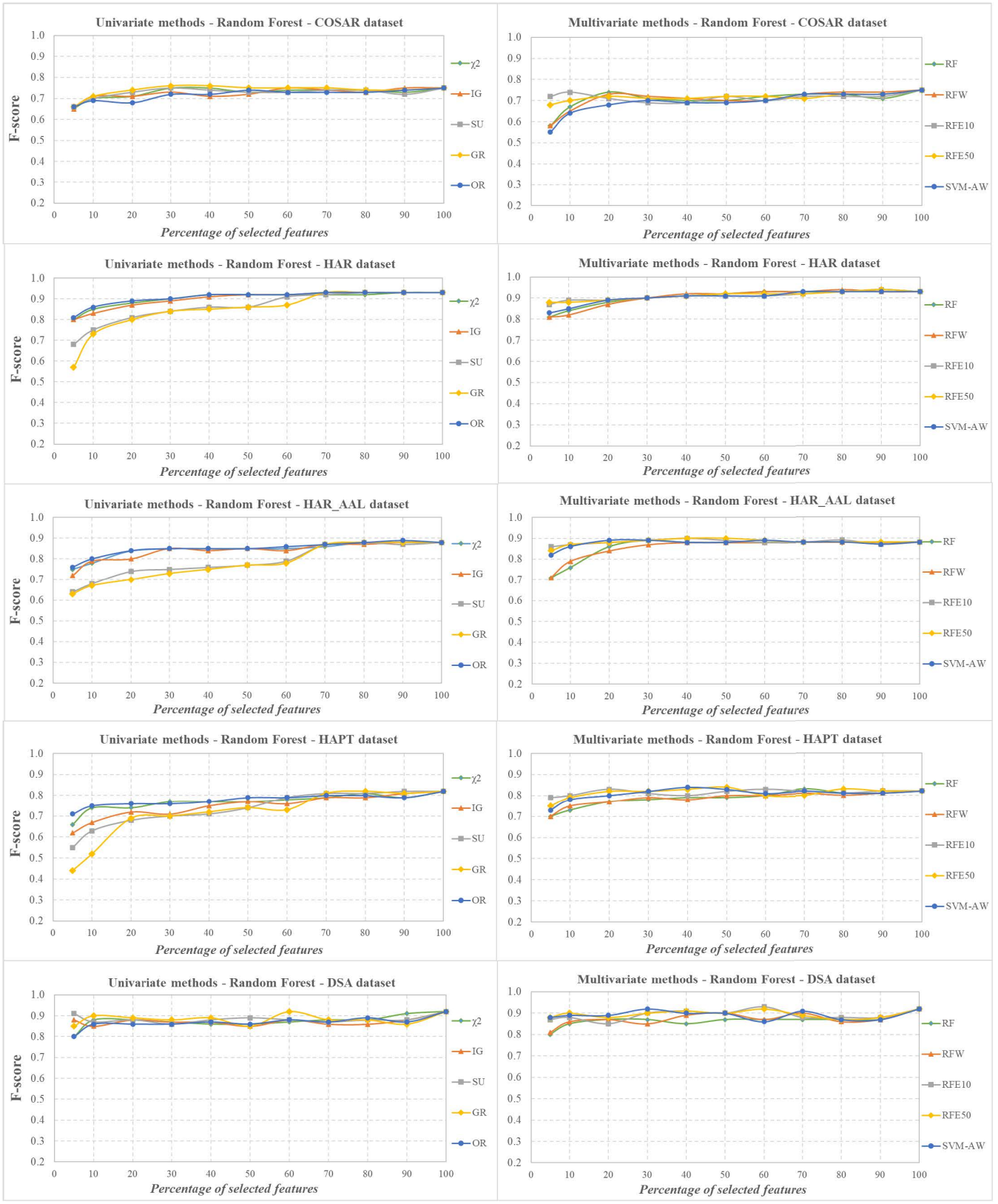

FIGURE 3. F-score performance of the *Random Forest* classifier, in conjunction with the univariate (on the left) and the multivariate (on the right) selection methods.

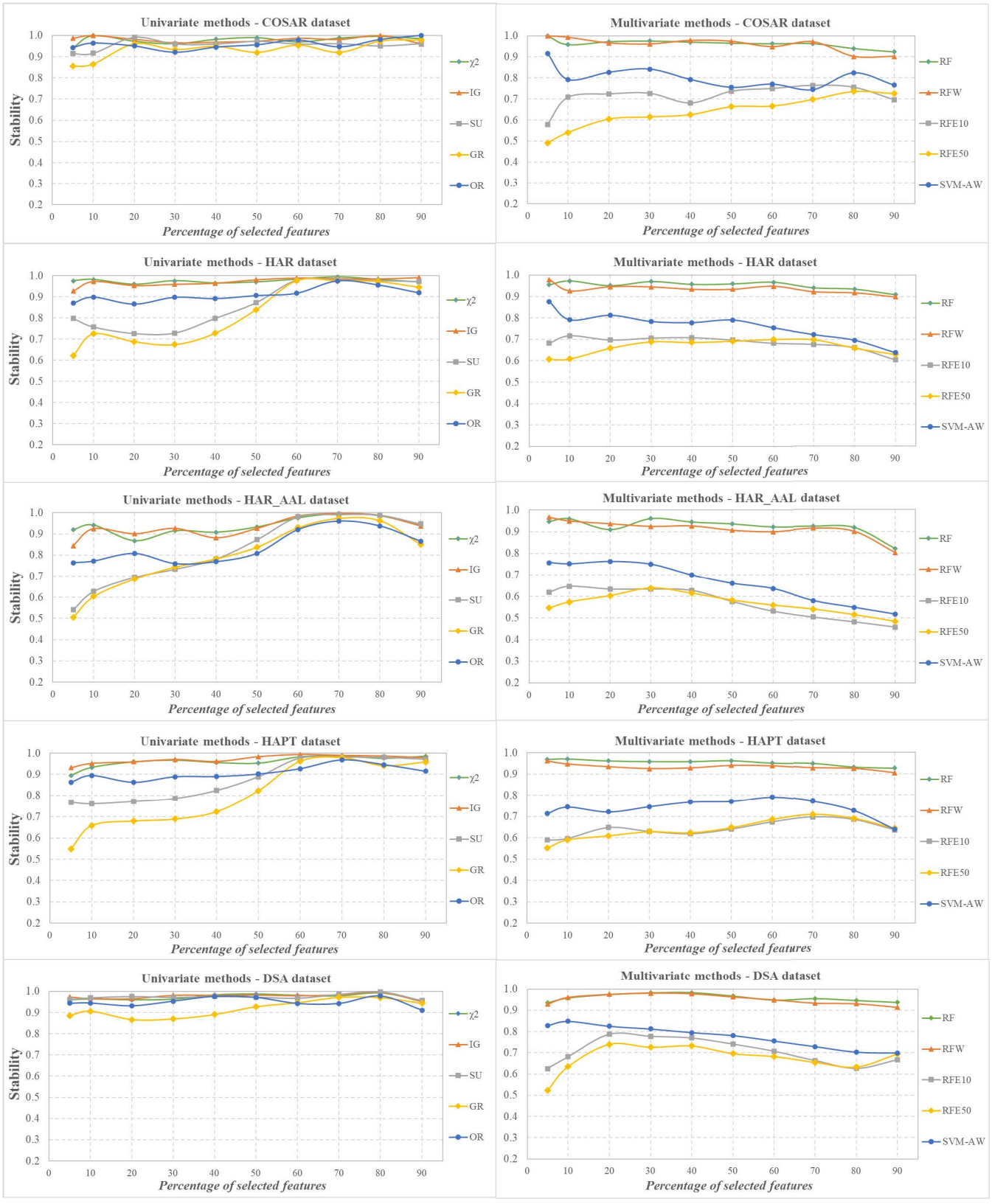

FIGURE 4. Stability trend for the univariate (on the left) and the multivariate (on the right) selection methods.

harmonic mean between the model *sensitivity* (i.e., the fraction of positive instances classified correctly) and the model *precision* (i.e., the fraction of correct predictions among all the instances assigned to the positive class), the F-score takes both the false positives and the false negatives into account, providing a reliable estimate of the model ability to recognize a given class (considered as positive). By measuring the average F-score across the different classes, we obtained an overall evaluation of the recognition performance of the induced models. For model induction, as anticipated in Section III-C, we employed both the *SMO* (Figure 2) and *Random Forest* (Figure 3) classifiers, in conjunction with ten different selection methods.

More in detail, for both the univariate (χ^2, *IG*, *SU*, *GR*, *OR*) and the multivariate (*RF*, *RFW*, *RFE10*, *RFE50*, *SVM-AW*) selection techniques, Figures 2 and 3 show the F-score trend for different percentages of selected features (the abscissa 100 corresponds to the model induced on the whole dataset, without any dimensionality reduction). In absolute terms, the *COSAR* dataset is the one where both *SMO* and *Random Forest* achieve the lowest recognition performance, due to the intrinsic difficulty of discriminating multiple activities using features extracted only from accelerometer signals. In the other four datasets, where we have a multi-sensor scenario (as detailed in Table 1), the average F-score is generally better, sometimes above 0.9, with quite similar trends for the two classifiers.

But the most interesting observation, for the purpose of our study, is that no significant degradation in performance is observed when the original dimensionality is reduced, regardless of the specific characteristics of the dataset at hand. In particular, 10-20% (or even less) of the original features may be sufficient, for some selection methods, to obtain recognition performances comparable to those achieved using the whole dataset. This reveals the appropriateness of introducing a suitable feature selection step into any activity recognition protocol in order to simplify the final models and make them more efficient.

When comparing the different selection techniques, the multivariate approaches, which can capture the interdependencies among the features, turn out to be overall more effective in terms of F-score. In particular, among the *SVM*-based methods, *RFE10* and *RFE50* sometimes have slightly superior performances but at a higher computational cost than the simpler *SVM-AW*, whose behavior is still satisfactory. As regards the *Relief*-based multivariate approaches (*RF* and *RFW*), quite good performance can be obtained with at most 20% of the features. Among the univariate methods, on the other hand, *SU* and *GR* overall exhibit the worst behavior, with an unsatisfactory performance for some datasets, especially when small feature subsets are selected. For the other univariate approaches, i.e., χ^2, *IG* and *OR*, feature subsets containing (at most) 20% of the features turn out to be sufficient to obtain quite good F-score values.

Overall, the strong potential of feature selection in this domain is witnessed by both Figures 2 and 3, despite some small differences between the curves reported for the *SMO* and *Random Forest* classifiers. However, as recognized by recent literature in the feature selection field, e.g., [31], [39], [61], [62], a good selection method should not only be effective (in terms of final predictive performance) but also as stable as possible to avoid that the selected subsets depend too much on the specific composition of the training data, thus becoming less useful in future applications of the model.

The results of the stability analysis we have performed on the five considered benchmarks (*COSAR*, *HAR*, *HAR_AAL*, *HAPT*, *DSA*) are shown in Figure 4, where the stability trend is reported for different percentages of selected features, according to the methodology detailed in sub-section III-B; specifically, such a methodology has been here implemented with the settings $K = 20$ and $f = 0.80$, which have proven suitable in similar studies [31], [39].

A first point to highlight regarding Figure 4 is that the differences observed in terms of stability are generally higher than those observed in terms of F-score (Figure 2 and Figure 3), revealing that some selection methods systematically exhibit a more stable behavior across the different datasets. In particular, χ^2 and *IG* turn out to be the most stable of the univariate approaches, followed by *OR*, while *SU* and *GR* have shown significantly lower stability in some datasets; *GR*, in particular, appears to be the least robust method in the univariate group. Among the multivariate approaches, on the other hand, the *Relief*-based methods (i.e., *RF* and *RFW*) have proven to be very stable, with similar trends for all levels of dimensionality reduction. Conversely, the *SVM*-based methods appear to be less robust, especially *RFE10* and *RFE50*, whose stability is always lower than the one of the simpler *SVM-AW*.

Overall, the univariate χ^2, *IG* and even *OR* show a quite good trade-off between F-score and stability and may therefore be an option to consider when inducing activity recognition models from mobile sensor datasets like those considered in our study. As well, the multivariate *Relief*-based approaches exhibit a good behavior when jointly considering both predictive performance and stability, at least for subsets containing more than 10% of the original features. It is worth mentioning, nevertheless, that the least stable methods could be made significantly more robust when implemented in a bootstrap-based ensemble version [39], thus envisaging margins of improvement for some selection techniques (but at the expense of the computational cost of the feature selection process).

Based on a wide set of experiments, the above results consolidate the findings of our preliminary study in this field [30], showing that feature selection can be effectively exploited to reduce the dimensionality of mobile sensor datasets, leading to robust and more efficient recognition models.

VI. DISCUSSION

The experimental analysis here presented complements and extends the findings of recent comparative studies

TABLE 2. *HAPT* dataset (1 accelerometer, 1 gyroscope, 561 features): different levels of dimensionality reduction and corresponding F-score performance.

Number of features (χ^2 feature selection)	*Features by sensor type*	*F-score (SMO classifier)*
29 (5%)	accelerometer: 29, gyroscope: 0	0.58
57 (10%)	accelerometer: 57, gyroscope: 0	0.82
85 (15%)	**accelerometer: 81, gyroscope: 4**	**0.85**
561 (100%)	accelerometer: 348, gyroscope: 213	0.86

TABLE 3. *DSA* dataset (5 accelerometers, 5 magnetometers, 5 gyroscopes, 1170 features): different levels of dimensionality reduction and corresponding F-score performance.

Number of features (χ^2 feature selection)	*Features by sensor type*	*F-score (SMO classifier)*
30 (2.5%)	accelerometer: 25, magnetometer: 5, gyroscope: 0	0.82
59 (5%)	accelerometer: 44, magnetometer: 15, gyroscope: 0	0.86
117 (10%)	**accelerometer: 71, magnetometer: 42, gyroscope:** 4	**0.89**
1170 (100%)	accelerometer: 390, magnetometer: 390, gyroscope: 390	0.89

conducted on similar activity recognition benchmarks [23], [29], [44], [45], providing stronger evidence of the suitability of the filter selection paradigm in this field.

Specifically, [23] discusses the advantages of leveraging classifier-independent selection approaches that can identify feature subsets conveying useful information for targeted populations and applications, regardless of the chosen classifier and the specific implementation settings. Compared to our work, the dataset considered in their study is relatively low-dimensional, with only seventy-six features, which allows efficiently using subsets-oriented filters (*CFS*, *FCBF*), along with a multivariate ranker (*ReliefF*), while the ranking-based approach is usually preferred in the presence of higher dimensionalities [44], [45]. Both subset-oriented filters (*CFS*, *MRMR*) and ranking-based filters (*Information Gain*, *Gain Ratio*, *Symmetrical Uncertainty*, *ReliefF*), along with wrapper and embedded methods, are also evaluated in [29], using a single sensor dataset involving 206 attributes (both time-domain and frequency-domain features), with empirical evidence of the effectiveness of the simpler univariate rankers in yielding good feature subsets for HAR models.

Through wider experiments on five high-dimensional benchmarks, our study has presented a more in-depth evaluation of the ranking-based selection approach, with both univariate and multivariate techniques, showing that they can be effectively employed in conjunction with different classifiers (see Figures 2 and 3). Actually, besides filters that inherently perform a classifier-independent selection, our experiments also encompass ranking methods that leverage some learning algorithm to compute the features' weights, as in the case of the *SVM*-based selectors. The final subsets produced by these methods, of the embedded category, are often used like those produced by filters, i.e., as a reduced feature space to train any potentially suitable classifier. However, as shown in Figure 4, this kind of ranker has proved to be overall less stable.

Our study has also shown that the adopted ranking-based approach allows to control the level of dimensionality reduction in a fine-grained way, in order to find the optimal trade-off between the number of the selected features and the resulting classification performance. Indeed, as discussed in section V, the original dimensionality can be reduced to a very great extent, with a final recognition performance similar to that achieved with the full feature set.

To provide concrete examples of such a dimensionality reduction, Tables 2 and 3 show the *SMO* classifier's F-score (averaged across the different classes, as in Figures 2 and 3) for the two multi-sensor benchmarks with the highest numbers of features/classes. Specifically, besides the full feature set, three feature subsets with smaller cardinality are considered, as selected by one of the univariate rankers that have shown the best tradeoff between predictive performance and stability, namely χ^2 (which also has a reduced computational cost compared to the multivariate ranking methods). As we can see, in the *HAPT* dataset (Table 2), only 15% of the original features are sufficient to obtain a final performance comparable to that achieved using the whole feature set, while 10% of the features turn out to be as predictive as the whole feature set in the *DSA* dataset (Table 3). Furthermore, interesting insight can be gained into the extent to which each sensor type contributes to the optimal set of features, with clear evidence of the high discriminative power of the features extracted from the accelerometer signals (that, however, are not sufficient to obtain the highest F-score values). This kind of analysis can leverage different ranking methods, as previously observed in Figures 2 and 3, and allows the design of fast-response recognition systems where only a reduced set of highly discriminative features need to be computed at operation time.

Overall, our results clearly show how beneficial feature selection can be in sensor-based activity recognition. However, some limitations exist in the current study that will be addressed in further investigations. Indeed, the explored ranking-based selection approach, although effective and generally more efficient than subset-oriented selection methods, may be sub-optimal in the presence of some degree of redundancy among the features. Hence, the potential impact of feature redundancy in this field should be better analyzed, for example by exploring hybrid selection strategies that first reduce the data dimensionality through a simple and efficient ranker and then further refine the resulting subset through

TABLE 4. Statistical measures used for feature extraction.

Measure	Formula	Datasets
Mean	$\bar{x} = \frac{\sum_{k=1}^{M} x(i)}{M}$	*COSAR* (14), *HAR* (46), *HAR_ALL* (46), *HAPT* (46), *DSA* (45)
Standard Deviation	$s = \sqrt{\frac{\sum_{i=1}^{M}(x(i)-\bar{x})^2}{M}}$	*HAR* (33), *HAR_ALL* (33), *HAPT* (33)
Variance	$s^2 = \frac{\sum_{i=1}^{M}(x(i)-\bar{x})^2}{M}$	*COSAR* (14), *DSA* (45)
Covariance	$Cov_{j,k} = \sum_{i=1}^{M}(x_j(i)-\bar{x}_j)(x_k(i)-\bar{x}_k)$	*COSAR* (18)
Correlation	$Cor_{j,k} = \frac{\sum_{i=1}^{M}(x_j(i)-\bar{x}_j)(x_k(i)-\bar{x}_k)}{\sqrt{\sum_{i=1}^{M}(x_j(i)-\bar{x}_j)^2(x_k(i)-\bar{x}_k)^2}}$	*COSAR* (18), *HAR* (15), *HAR_ALL* (15), *HAPT* (15)
Median	$m = \begin{cases} \tilde{x}(\frac{M+1}{2}) \text{ if } M \text{ is odd} \\ \frac{\tilde{x}(\frac{M}{2})+\tilde{x}(\frac{M}{2}+1)}{2} \text{ if } M \text{ is even} \end{cases}$	*COSAR* (14), *HAR* (33), *HAR_ALL* (33), *HAPT* (33)
Min/Max Values	$\min = \min_{i=1,\ldots,M}\{x(i)\}$; $\max = \max_{i=1,\ldots,M}\{x(i)\}$	*COSAR* (40), *HAR* (79), *HAR_ALL* (79), *HAPT* (79), *DSA* (90)
90th Percentile	$p_{90} = \tilde{x}(0.90 \cdot M)$	*COSAR* (14)
Harmonic Mean	$h = \frac{M}{\sum_{i=1}^{M}\frac{1}{x(i)}}$	*COSAR* (2)
Signal Energy	$SE_j = \sum_{i=1}^{M}\lvert X_j(i)\rvert^2$	*HAR* (159), *HAR_ALL* (159), *HAPT* (159)
Entropy	$E_j = -\sum_{i=1}^{M}\frac{\lvert X_j(i)\rvert^2}{\sum_{k=1}^{M}\lvert X_j(k)\rvert^2}\log_2\left(\frac{\lvert X_j(i)\rvert^2}{\sum_{k=1}^{M}\lvert X_j(k)\rvert^2}\right)$	*HAR* (33), *HAR_ALL* (33), *HAPT* (33)
Skewness	$Sk = \frac{\sum_{i=1}^{M}\frac{1}{M}(x(i)-\bar{x})^3}{\sqrt{\left(\sum_{i=1}^{M}\frac{1}{M}(x(i)-\bar{x})^2\right)^3}}$	*HAR* (13), *HAR_ALL* (13), *HAPT* (13), *DSA* (45)
Kurtosis	$Kur = \frac{\sum_{i=1}^{M}\frac{1}{M}(x(i)-\bar{x})^4}{\left(\sum_{i=1}^{M}\frac{1}{M}(x(i)-\bar{x})^2\right)^2} - 3$	*COSAR* (14), *HAR* (13), *HAR_ALL* (13), *HAPT* (13), *DSA* (45)
Autoregressive Coefficients	If $x_j(i) = \sum_{i=1}^{p}\phi_k x_j(i-k) + \epsilon(i)$ is the Autoregressive Model, then ϕ_k are the **Autoregressive Coefficients**	*HAR* (80), *HAR_ALL* (80), *HAPT* (80)
Interquartile Range	$IqR = Q_3 - Q_1$, Q_1 and Q_3 are the first and third quartile.	*HAR* (33), *HAR_ALL* (33), *HAPT* (33)
Signal Magnitude Area	$SMA = \frac{1}{M}\sum_{i=1}^{M}\lvert x_1(i)\rvert + \lvert x_2(i)\rvert + \lvert x_3(i)\rvert$	*HAR* (17), *HAR_ALL* (17), *HAPT* (17)
Angle	$A(i) = \arccos\frac{x_3(i)}{\sqrt{x_1(i)^2+x_2(i)^2+x_3(i)^2}}$	*HAR* (7), *HAR_ALL* (7), *HAPT* (7)
Autocorrelation	$Auto(k) = \frac{1}{M-k}\sum_{i=0}^{M-k-1}(x(i)-\bar{x})(x(i-k)-\bar{x})$	*DSA* (450)
Peaks	Let be $S_{DFT}(k) = \sum_{i=1}^{M}x(i)e^{-j2\pi ki/M}$ the k-th element of the 1-D M-point Discrete Fourier Transform, then **Peaks** are the maximum five Fourier peaks.	*DSA* (225)
Frequency Peaks	The frequency values that correspond to the Fourier **Peaks**	*DSA* (225)

a more sophisticated search strategy that can remove highly correlated features [38]. But this would increase the computational cost of the selection process, requiring careful cost-benefit analysis. Although some recent works have applied a hybrid selection approach in the HAR field [24], [49], [54], there is a lack of comparative studies that investigate the effects of different hybrid strategies in terms of final recognition performance as well as selection stability and computational efficiency.

VII. CONCLUDING REMARKS

Recent advances in mobile sensor technology have opened up growing opportunities for HAR applications in a variety of fields, from personal wellness to health monitoring. However, as data collection becomes easier and easier, fully exploiting the available data to build reliable models for activity recognition remains challenging. Further, when dealing with wearable devices, the issues of limited resources and energy consumption cannot be neglected, posing additional requirements in terms of efficiency of the induced models. In such a perspective, feature selection may have an important role since it can significantly reduce the data dimensionality by removing irrelevant and noisy features.

In this paper, we have explored the potential of feature selection in the HAR field, leveraging five public mobile sensor datasets with heterogeneous characteristics and relying on a methodological framework that considers both the predictive power and the stability of the selected features subsets (i.e., their robustness to perturbations in the input data). Ten different selection methods have been comparatively evaluated, revealing the suitability of univariate ranking approaches, such as *Chi Squared* and *Information Gain*, as well as of some multivariate approaches, such as *ReliefF*, which exhibit a quite good trade-off between recognition performance and selection stability. Encompassing different levels of dimensionality reduction, our analysis has shown that the original number of features can be reduced to a very large extent in all the considered benchmarks, without any worsening of performance.

Such a large comparative study gives evidence of the potential benefits of systematically exploiting feature selection when inducing HAR models, also providing methodological insights into how to evaluate the robustness of the selection outcome and choose the optimal level of dimensionality reduction. This line of research is worthy of further investigation, given the lack of such kind of comparative studies in this field.

As future work, our study will be enlarged to better investigate the extent to which the behavior and the stability pattern of a given selection algorithm may depend on the underlying feature extraction strategy (i.e., the type and number of features extracted for each sensor). Furthermore, the ranking-based selection framework here adopted will be compared with different and more sophisticated methodological approaches, including hybrid techniques that exploit different heuristics at different stages of the selection process and ensemble techniques that suitably combine the outcome of different selectors.

APPENDIX I. FEATURE DESCRIPTION

In this appendix, our aim is to give an overview of the statistical measures that were computed to build the feature vectors of the considered datasets (i.e., *COSAR*, *HAR*, *HAR_AAL*, *HAPT*, *DSA*).

All the measures are listed in Table 4, with the corresponding formula. Note that the third column of the table cites the datasets that use the measure and, in brackets, the number of times it was calculated.

In each formula of the table, the variable x represents the signal. Depending on the dataset used and the feature calculated, this signal x can be understood as an acceleration, an inclination, an angular velocity, a magnitude signal, or a Jerk signal (that is the first time derivative of the acceleration).

As mentioned in Section IV, the signals were segmented into sliding windows for feature extraction; specifically, all the measures reported in Table 4 are calculated on windows containing a number of observations denoted by M.

It should also be considered that the sensors used in the experiments (accelerometers, magnetometers, and gyroscopes) acquire the three spatial components of each signal; thus, in some measures, x can represent the component on the X, Y or Z axes of the considered signal. Other features, such as correlation and covariance, are calculated using two or more components on the axes; therefore, in the formulas it has been necessary to differentiate these components, indicating with x_1, x_2 and x_3 the components on the X, Y and Z axes respectively. Some other features, such as median and percentiles, are calculated using the window signal values sorted in ascending order, so we have indicated the ordered values differently using $\tilde{x}$.

As already pointed out in Section IV, the feature vectors were computed in both the time domain and the frequency domain; some measures in fact, e.g. Entropy, have been calculated in the frequency domain; consequently, in Table 4, the frequency signal is indicated with the capital letter X, to distinguish it from the signal x in the time domain.

REFERENCES

[1] E. De-La-Hoz-Franco, P. Ariza-Colpas, J. M. Quero, and M. Espinilla, "Sensor-based datasets for human activity recognition—A systematic review of literature," *IEEE Access*, vol. 6, pp. 59192–59210, 2018.

[2] G. Cicirelli, R. Marani, A. Petitti, A. Milella, and T. D'Orazio, "Ambient assisted living: A review of technologies, methodologies and future perspectives for healthy aging of population," *Sensors*, vol. 21, no. 10, p. 3549, May 2021.

[3] D. Hendry, K. Chai, A. Campbell, L. Hopper, P. O'Sullivan, and L. Straker, "Development of a human activity recognition system for ballet tasks," *Sports Med.*, vol. 6, no. 1, pp. 1–10, Dec. 2020.

[4] T. Steels, B. Van Herbruggen, J. Fontaine, T. De Pessemier, D. Plets, and E. De Poorter, "Badminton activity recognition using accelerometer data," *Sensors*, vol. 20, no. 17, p. 4685, Aug. 2020.

[5] J. Qi, P. Yang, A. Waraich, Z. Deng, Y. Zhao, and Y. Yang, "Examining sensor-based physical activity recognition and monitoring for healthcare using Internet of Things: A systematic review," *J. Biomed. Inform.*, vol. 87, pp. 138–153, Nov. 2018.

[6] K. Barbaro, "Automated sensing of daily activity: A new lens into development," *Develop. Psychobiol.*, vol. 61, no. 3, pp. 444–464, Apr. 2019.

[7] E. Khodabandehloo, D. Riboni, and A. Alimohammadi, "HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline," *Future Gener. Comput. Syst.*, vol. 116, pp. 168–189, Mar. 2021.

[8] W. S. Lima, E. Souto, K. El-Khatib, R. Jalali, and J. Gama, "Human activity recognition using inertial sensors in a smartphone: An overview," *Sensors*, vol. 19, no. 14, p. 3213, Jul. 2019.

[9] F. Serpush, M. B. Menhaj, B. Masoumi, and B. Karasfi, "Wearable sensor-based human activity recognition in the smart healthcare system," *Comput. Intell. Neurosci.*, vol. 2022, Feb. 2022, Art. no. 1391906.

[10] M. Straczkiewicz, P. James, and J.-P. Onnela, "A systematic review of smartphone-based human activity recognition methods for health research," *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–15, Dec. 2021.

[11] F. Demrozi, G. Pravadelli, A. Bihorac, and P. Rashidi, "Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey," *IEEE Access*, vol. 8, pp. 210816–210836, 2020.

[12] B. Nguyen, Y. Coelho, T. Bastos, and S. Krishnan, "Trends in human activity recognition with focus on machine learning and power requirements," *Mach. Learn. Appl.*, vol. 5, Sep. 2021, Art. no. 100072.

[13] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explor. Newslett.*, vol. 12, no. 2, pp. 74–82, May 2011.

[14] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomputing*, vol. 171, pp. 754–767, Jan. 2016.

[15] M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," *Future Gener. Comput. Syst.*, vol. 81, pp. 307–313, Apr. 2018.

[16] M. Ehatisham-Ul-Haq, M. A. Azam, Y. Asim, Y. Amin, U. Naeem, and A. Khalid, "Using smartphone accelerometer for human physical activity and context recognition in-the-wild," *Proc. Comput. Sci.*, vol. 177, pp. 24–31, Jan. 2020.

[17] J. Sena, J. Barreto, C. Caetano, G. Cramer, and W. R. Schwartz, "Human activity recognition based on smartphone and wearable sensors using multiscale DCNN ensemble," *Neurocomputing*, vol. 444, pp. 226–243, Jul. 2021.

[18] E. J. Huang, K. Yan, and J.-P. Onnela, "Smartphone-based activity recognition using multistream movelets combining accelerometer and gyroscope data," *Sensors*, vol. 22, no. 7, p. 2618, Mar. 2022.

[19] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1192–1209, Nov. 2013.

[20] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, "Trends in human activity recognition using smartphones," *J. Reliable Intell. Environ.*, vol. 7, pp. 189–213, Sep. 2021.

[21] H. F. Nweke, Y. W. Teh, M. A. Al-Garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," *Expert Syst. Appl.*, vol. 105, pp. 233–261, Sep. 2018.

[22] P. Gupta and T. Dallas, "Feature selection and activity recognition system using a single triaxial accelerometer," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1780–1786, Jun. 2014.

[23] N. A. Capela, E. D. Lemaire, and N. Baddour, "Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients," *PLoS ONE*, vol. 10, no. 4, Apr. 2015, Art. no. e0124414.

[24] N. Ahmed, J. I. Rafiq, and M. R. Islam, "Enhanced human activity recognition based on smartphone sensor data using hybrid feature selection model," *Sensors*, vol. 20, no. 1, p. 317, Jan. 2020.

[25] A. Aguileta, R. Brena, O. Mayora, E. Molino-Minero-Re, and L. Trejo, "Multi-sensor fusion for activity recognition—A survey," *Sensors*, vol. 19, p. 3808, Jan. 2019.

[26] A. Reiss, G. Hendeby, and D. Stricker, "A competitive approach for human activity recognition on smartphones," in *Proc. Eur Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn (ESANN)*, Bruges, Belgium, 2013, pp. 460–4550.

[27] K. K. A. Rahim, I. Elamvazuthi, L. Izhar, and G. Capi, "Classification of human daily activities using ensemble methods based on smartphone inertial sensors," *Sensors*, vol. 18, no. 12, p. 4132, Nov. 2018.

[28] R. Zhu, Z. Xiao, Y. Li, M. Yang, Y. Tan, L. Zhou, S. Lin, and H. Wen, "Efficient human activity recognition solving the confusing activities via deep ensemble learning," *IEEE Access*, vol. 7, pp. 75490–75499, 2019.

[29] J. Chong, P. Tjurin, M. Niemelä, T. Jämsä, and V. Farrahi, "Machine-learning models for activity class prediction: A comparative study of feature selection and classification algorithms," *Gait Posture*, vol. 89, pp. 45–53, Sep. 2021.

[30] A. Loddo, B. Pes, and D. Riboni, "Feature selection in mobile activity recognition: A comparative study," in *Proc. 22nd IEEE Int. Conf. Mobile Data Manage. (MDM)*, Jun. 2021, pp. 181–186.

[31] B. Pes, "Feature selection for high-dimensional data: The issue of stability," in *Proc. IEEE 26th Int. Conf. Enabling Technol., Infrastructure Collaborative Enterprises (WETICE)*, Jun. 2017, pp. 170–175.

[32] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, pp. 1060–1073, Jun. 2022.

[33] *COSAR Activity Recognition Dataset*. Accessed: May 2022. [Online]. Available: https://everywarelab.di.unimi.it/index.php/palspot

[34] *Human Activity Recognition Using Smartphones Dataset*. Accessed: May 2022. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition +Using+Smartphones

[35] *Smartphone Dataset for Human Activity Recognition in Ambient Assisted Living*. Accessed: May 2022. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Smartphone+Dataset+for+Human+Activity+Recognition+(HAR)+in+Ambient+Assisted+Living+(AAL)

[36] *Smartphone-Based Recognition of Human Activities and Postural Transitions Dataset*. Accessed: May 2022. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/smartphone-based+recognition+of+human+activities+and+postural+transitions

[37] *Daily Sports Activities Dataset*. Accessed: May 2022. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities

[38] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, *Feature Selection for High-Dimensional Data* (Artificial Intelligence: Foundations, Theory, and Algorithms). Cham, Switzerland: Springer, 2015.

[39] B. Pes, "Ensemble feature selection for high-dimensional data: A stability analysis across multiple domains," *Neural Comput. Appl.*, vol. 32, no. 10, pp. 5951–5973, May 2020.

[40] L. C. Jatoba, U. Grossmann, C. Kunze, J. Ottenbacher, and W. Stork, "Context-aware mobile health monitoring: Evaluation of different pattern recognition methods for classification of physical activity," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Vancouver, BC, Canada, Aug. 2008, pp. 5250–5253.

[41] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher, "Activity recognition and monitoring using multiple sensors on different body positions," in *Proc. Int. Workshop Wearable Implant. Body Sensor Netw. (BSN)*, Cambridge, MA, USA, 2006, pp. 113–116.

[42] Z. Wang, D. Wu, J. Chen, A. Ghoneim, and M. A. Hossain, "A triaxial accelerometer-based human activity recognition via EEMD-based features and game-theory-based feature selection," *IEEE Sensors J.*, vol. 16, no. 9, pp. 3198–3207, May 2016.

[43] G. Chetty, M. White, and F. Akther, "Smart phone based data mining for human activity recognition," *Proc. Comput. Sci.*, vol. 46, pp. 1181–1187, Jan. 2015.

[44] L. Chen, S. Fan, V. Kumar, and Y. Jia, "A method of human activity recognition in transitional period," *Information*, vol. 11, no. 9, p. 416, Aug. 2020.

[45] S. Fan, Y. Jia, and C. Jia, "A feature selection and classification method for activity recognition based on an inertial sensing unit," *Information*, vol. 10, no. 10, p. 290, Sep. 2019.

[46] I. Amezzane, Y. Fakhri, M. El Aroussi, and M. Bakhouya, "Analysis and effect of feature selection over smartphone-based dataset for human activity recognition," in *Emerging Technologies for Developing Countries*. Cham, Switzerland: Springer, 2018, pp. 214–219.

[47] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 1, pp. 1–26, Dec. 2020.

[48] S. K. Bashar, A. Al Fahim, and K. H. Chon, "Smartphone based human activity recognition with feature selection and dense neural network," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Montreal, QC, Canada, Jul. 2020, pp. 5888–5891.

[49] H. F. Nweke, Y. W. Teh, G. Mujtaba, U. R. Alo, and M. A. Al-Garadi, "Multi-sensor fusion based on multiple classifier systems for human activity identification," *Hum.-Centric Comput. Inf. Sci.*, vol. 9, no. 1, pp. 1–44, Dec. 2019.

[50] M. Zhang and A. Sawchuk, "A feature selection-based framework for human activity recognition using wearable multimodal sensors," in *Proc. 6th Int. ICST Conf. Body Area Netw.*, Brussels, Belgium, 2011, pp. 92–98.

[51] J. Suto, S. Oniga, and P. P. Sitar, "Comparison of wrapper and filter feature selection algorithms on human activity recognition," in *Proc. 6th Int. Conf. Comput. Commun. Control (ICCCC)*, Oradea, Romania, May 2016, pp. 124–129.
[52] L. Brezočnik, I. Fister, and V. Podgorelec, "Swarm intelligence algorithms for feature selection: A review," *Appl. Sci.*, vol. 8, no. 9, p. 1521, Sep. 2018.
[53] A. M. Helmi, M. A. A. Al-Qaness, A. Dahou, R. Damaševičius, T. Krilavičius, and M. A. Elaziz, "A novel hybrid gradient-based optimizer and grey wolf optimizer feature selection method for human activity recognition using smartphone sensors," *Entropy*, vol. 23, no. 8, p. 1065, Aug. 2021.
[54] Y. Tian, J. Zhang, L. Li, and Z. Liu, "A novel sensor-based human activity recognition method based on hybrid feature selection and combinational optimization," *IEEE Access*, vol. 9, pp. 107235–107249, 2021.
[55] C. Fan and F. Gao, "Enhanced human activity recognition using wearable sensors via a hybrid feature selection method," *Sensors*, vol. 21, no. 19, p. 6434, Sep. 2021.
[56] V. Bolón-Canedo, A. Alonso-Betanzos, L. Morán-Fernández, and B. Cancela, "Feature selection: From the past to the future," in *Advances in Selected Artificial Intelligence Areas, Learning and Analytics in Intelligent Systems*, vol. 24. Cham, Switzerland: Springer, 2022.
[57] 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.
[58] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, Jan. 2018.
[59] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," *Inf. Fusion*, vol. 52, pp. 1–12, Dec. 2019.
[60] N. Almugren 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.
[61] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," *J. Mach. Learn. Res.*, vol. 18, no. 174, pp. 1–54, 2018.
[62] 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)*, Las Vegas, NV, USA, Aug. 2012, pp. 356–363.
[63] L. Kuncheva, "A stability index for feature selection," in *Proc. 25th IASTED Int. Multi-Conf., Artif. Intell. Appl. (AIAP)*, 2007, pp. 390–395.
[64] B. Pes and G. Lai, "Cost-sensitive learning strategies for high-dimensional and imbalanced data: A comparative study," *PeerJ Comput. Sci.*, vol. 7, p. e832, Dec. 2021.
[65] P. N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*, 2nd ed. New York, NY, USA: Pearson, 2019.
[66] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *DATA MINING: Practical Machine Learning Tools and Techniques*. Cambridge, MA, USA: Morgan Kaufmann, 2016.
[67] R. Holte, "Very simple classification rules perform well on most commonly used datasets," *Mach. Learn.*, vol. 11, pp. 63–91, Apr. 1993.
[68] 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.
[69] A. Rakotomamonjy, "Variable selection using SVM based criteria," *J. Mach. Learn. Res.*, vol. 3, pp. 1357–1370, Mar. 2003.
[70] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Recent advances and emerging challenges of feature selection in the context of big data," *Knowl.-Based Syst.*, vol. 86, pp. 33–45, Sep. 2015.
[71] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, *Feature Extraction: Foundations and Applications* (Studies in Fuzziness and Soft Computing). Berlin, Germany: Springer-Verlag, 2006.
[72] *Weka 3: Data Mining Software in Java*. Accessed: May 2022. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/
[73] 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, 2001.
[74] B. Pes, "Learning from high-dimensional and class-imbalanced datasets using random forests," *Information*, vol. 12, no. 8, p. 286, Jul. 2021.
[75] D. Riboni and C. Bettini, "COSAR: Hybrid reasoning for context-aware activity recognition," *Pers. Ubiquitous Comput.*, vol. 15, no. 3, pp. 271–289, Mar. 2011.
[76] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. ESANN*, Bruges, Belgium, 2013, pp. 437–442.
[77] K. Davis, E. Owusu, V. Bastani, L. Marcenaro, J. Hu, C. Regazzoni, and L. Feijs, "Activity recognition based on inertial sensors for ambient assisted living," in *Proc. 19th Int. Conf. Inf. Fusion (FUSION)*, Heidelberg, Germany, 2016, pp. 371–378.
[78] K. Altun, B. Barshan, and O. Tunçel, "Comparative study on classifying human activities with miniature inertial and magnetic sensors," *Pattern Recognit.*, vol. 43, no. 10, pp. 3605–3620, Oct. 2010.
[79] B. Barshan and M. C. Yüksek, "Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units," *Comput. J.*, vol. 57, no. 11, pp. 1649–1667, 2014.

MARCO MANOLO MANCA received the master's degree in mathematics from the University of Cagliari, Italy, in 2019, where he is currently pursuing the Ph.D. degree in computer science.

He is with the CRS4 Research Center. His current research interests include machine learning applied to signal analysis and smart energy systems.

BARBARA PES (Member, IEEE) is a Researcher (permanent position) with the Department of Mathematics and Computer Science, University of Cagliari, Italy, where she taught/teaches 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 bioinformatics. She is the author/coauthor of more than 80 papers published in international conferences, books, and journals. Currently, her main research interests include the fields of data mining and machine learning, classification of high-dimensional data, and advanced feature selection methods.

DANIELE RIBONI (Member, IEEE) is an Associate Professor of computer science with the University of Cagliari. His research interests include context-awareness, activity recognition, knowledge management, and privacy issues in pervasive and mobile computing. He served as TPC Chair and TPC Vice Chair for different conferences and workshops in the field, including IEEE Pervasive Computing and Communication (PerCom) and the International Conference on Intelligent Environments (IE). His contributions appear in major conferences and journals.