Random Forest Dissimilarity Based Multi-View Learning for Radiomics Application

Hongliu CAO, Simon BERNARD, Robert SABOURIN, Laurent HEUTTE

PII: S0031-3203(18)30400-X
DOI: https://doi.org/10.1016/j.patcog.2018.11.011
Reference: PR 6708

To appear in: *Pattern Recognition*

Received date: 21 February 2018
Revised date: 28 September 2018
Accepted date: 16 November 2018

Please cite this article as: Hongliu CAO, Simon BERNARD, Robert SABOURIN, Laurent HEUTTE, Random Forest Dissimilarity Based Multi-View Learning for Radiomics Application, *Pattern Recognition* (2018), doi: https://doi.org/10.1016/j.patcog.2018.11.011

Highlights

- Propose a Random forest dissimilarity based method for multi-view learning
- Study the effect of hyperparameters on the quality of random forest dissimilarity
- Compare the proposed method to the state of art Radiomics solutions
- Compare the proposed method to multi-view learning approaches
- Show that the proposed approach outperforms the state-of-the-art methods

Random Forest Dissimilarity Based Multi-View Learning for Radiomics Application

Hongliu CAO*[a,b], Simon BERNARD[a], Robert SABOURIN[b], Laurent HEUTTE[a]

[a] *Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France*
[b] *Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle, École de Technologie Supérieure, Université du Québec, Montreal, Canada*

Abstract

Radiomics is a medical imaging technique that aims at extracting a large amount of features from one or several modalities of medical images, in order to help diagnose and treat diseases like cancers. Many recent studies have shown that Radiomics features can offer a lot of useful information that physicians cannot extract from these images, and can be efficiently associated with other information like gene or protein data. However, most of the classification studies in Radiomics report the use of feature selection methods without identifying the underlying machine learning challenges. In this paper, we first show that the Radiomics classification problem should be viewed as a high dimensional, low sample size, multi-view learning problem. Then, we propose a dissimilarity-based method for merging the information from the different views, based on Random Forest classifiers. The proposed approach is compared to different state-of-the-art Radiomics and multi-view solutions, on different public multi-view datasets as well as on Radiomics datasets. In particular, our experiments show that the proposed approach works better than the state-of-the-art methods from the Radiomics, as well as from the multi-view learning literature.

Keywords: Radiomics, dissimilarity space, random forest, machine learning, feature selection, multi-view learning, high dimension, low sample size.

*Corresponding author
Email addresses: caohongliu@gmail.com (Hongliu CAO), simon.bernard@univ-rouen.fr (Simon BERNARD), Robert.Sabourin@etsmtl.ca (Robert SABOURIN), laurent.heutte@univ-rouen.fr (Laurent HEUTTE)

1. Introduction

Radiomics is a medical imaging technique that has aroused great interest over the past few years [1]. The core principle of Radiomics is to extract a very large number of features from multiple medical imaging modalities, in order to help to automate and improve diagnosis and treatment of certain diseases, like cancer for example. Radiomic features, along with the use of machine learning techniques, allow to retrieve useful information from images, usually invisible to the naked eye [2]. Furthermore, the information is thought nowadays to be complementary to clinical, pathological, and genomic information [3, 4].

From our point of view, the main reason behind the success of Radiomics is that models and classifiers are learned from many different families of features, which vehicle different types of information that assume to be complementary to each other. This can stem from the use of different extractors applied on the same image modality, or from the joint use of different image and/or non-image modalities [5, 6, 7, 8, 9]. In the case of cancer treatment for example, the use of several sets of features, with different clinical interpretations, efficiently helps to capture the important heterogeneity of cancers [10, 11]. The Radiomics features for cancer treatment, not only reflect different biologic mechanisms, such as gene-expression patterns or cell cycling pathways [12], but also provide more valuable clinical information than conventional medical imaging techniques [13].

From the machine learning point of view, learning from Radiomic features can be characterized through three challenges:

1. **Small sample size**: Due to different acquisition protocols, different laws or politic issues, Radiomics datasets are usually very small, with often no more than 50 patients [14, 15, 16, 17]. It is a recurrent difficulty in medical pattern recognition tasks, that is exacerbated here by the fact that each institution is using its own medical protocol, based on its own image acquisition parametrization.
2. **High dimensional feature space**: By definition, a large amount of Radiomic features has to be extracted from the images. High dimensional feature spaces are also usual in medical imaging, but it is again accentuated here by the fact that several

families of features are needed for tackling the strong heterogeneity of cancers. No quantitative definition of 'large amount' is given in Radiomics studies but the number of features is usually above 4 or 5 times the number of learning instances [5, 18, 7]. For example, the Radiomics dataset used in [19] is made up of 84 patients and 6746 features.

3. **Multiple sets of features**: In many studies, the Radiomics features describe the tumor intensity, shape and texture [5], but also sometimes come from other modalities like clinical or genomic information. Exploiting the complementarity of the different sets of features is a challenging task that requires dedicated machine learning techniques.

The first two challenges form what is called an HDLSS (High Dimensional Low Sample Size) machine learning problem, which is common in medical pattern recognition problems. The third challenge, however, is more specific to Radiomics and can be viewed as a multi-view learning problem. Most of the Radiomics works in the literature ignore this specificity and treat Radiomics as a single-view problem by concatenating all the Radiomics features to form a unique, very high dimensional feature space. As a consequence, most of these works use some feature selection techniques to find a relevant feature subset while reducing the dimension at the same time. Usually in that case, the most used feature selection methods are filter methods that select a subset of features independently from the learning phase [7, 20, 1]. The main reason in our opinion for using filter methods in that case is that taking the classifier outcome into account would require to use a validation set, which is difficult in a HDLSS setting [13, 6, 5]. However, when few features are selected regardless the learning task, a lot of useful information may be lost, that could have been important though to efficiently tackle this task.

In this work, we propose to consider the Radiomics learning task as an HDLSS multi-view learning problem. In the multi-view learning literature [21], three main approaches are typically proposed: (i) early integration methods that concatenate the views together and treat them as a single-view problem, like in the Radiomics works mentioned above;

(ii) intermediate integration methods that fuse the views before learning; and (iii) late integration methods that combine models learned separately on each view.

In a preliminary study [22], we have shown the potential of multi-view approaches, i.e. intermediate and late integration approaches, in comparison with early integration methods. We have also shown the interest of using the RFD (Random Forest Dissimilarity) mechanism for tackling the HDLSS challenges, a mechanism that takes both feature and class information into consideration, and that efficiently deals with high dimensional data without feature decimation. The present work is an extended version of [22] by adding: (i) the theoretical justification of the two proposed methods RFSVM and RFDIS based on the RFD measure; (ii) a deep analysis of the influence of the main hyperparameters on the quality of the RFD measure along with recommendations for parameterization; (iii) an extended experimental validation on more multi-view datasets, with comparison to more state-of-the-art methods. Notably, the two proposed approaches are compared to a similar dissimilarity-based learning method from the literature (named MDSRF in the following) and a Multiple Kernel Learning (EasyMKL) method that offers an alternate way to combine multiple dissimilarity-based representations.

The remainder of this paper is organized as follows. The related works in Radiomics applications and multi-view learning are discussed in Section 2. In Section 3, the dissimilarity-based representation is introduced and the parametrization of the RFD measure is studied. In Section 4, two dissimilarity-based multi-view learning solutions are proposed. The protocol of our experiments and the results are described in Section 5. The final conclusion and future works are given in Section 6.

2. Related Works

As explained in the introduction, Radiomics can be seen as an HDLSS multi-view learning problem. In this section, we give an overview of the different multi-view learning approaches that have been proposed in the literature and analyze if they are suitable for HDLSS problems like Radiomics. According to [23], there are three main kinds of multi-view approaches:

early integration, intermediate integration and late integration. Their underlying principles are detailed below and illustrated in Figure 1.

1. Early integration methods directly concatenate different views together and adopt a traditional single-view learning approach [23]. Most works in Radiomics belong to this category of methods.
2. Late integration methods firstly build separate models on each view and combine them afterward. Co-training and MCS (Multiple Classifier Systems) are the two main solutions of this category.
3. Intermediate integration methods fuse the information from different views at the feature level and perform learning in the joint feature space [24, 21]. The most used feature integration methods include subspace learning and MRF (Multi-Representation Fusion) methods [21].

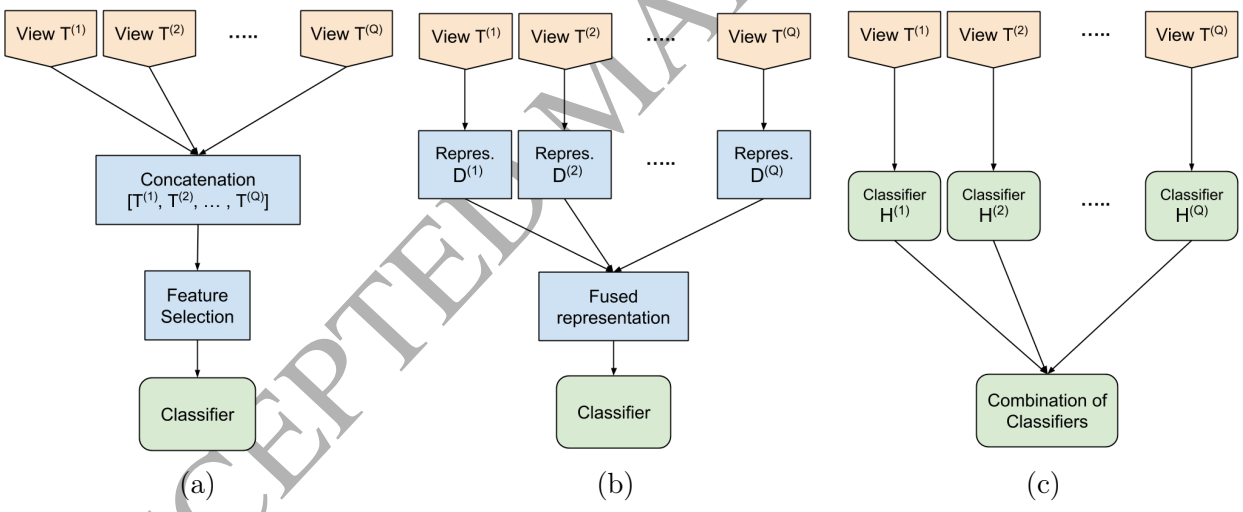


Figure 1: Flowcharts of the three multi-view learning approaches with (a) early integration, (b) intermediate integration and (c) late integration

2.1. Early integration methods in Radiomics

Most state-of-the-art methods in Radiomics concatenate all the Radiomics features together and apply a feature selection procedure (Figure 1 (a)). These feature selection methods are mostly used to reduce the redundancy, noise, or irrelevant features without an

expected significant loss of performance. Generally speaking, feature selection methods can be divided into three groups: filter methods, wrapper methods, and embedded methods [25, 26, 27].

Filter methods consist in ranking features according to a given criterion measured on each feature separately, and in using only the best ones according to a predefined number or to a threshold [25]. Several studies have compared different criteria for filter methods, along with reliable Machine Learning methods, to find the optimal combination for Radiomics applications [7, 6, 9]. From these studies, the most efficient criteria are WLCX (Wilcoxon), MRMR (minimum redundancy maximum relevance), RELF (Relief) and MIFS (mutual information feature selection) [18, 8, 9, 6]. In [7], 14 criteria and 12 different classifiers are compared on images of lung cancer patients. The best results are obtained using WLCX, MRMR and MIFS. In [9], 24 criteria along with 3 classifiers are compared, the best performance being obtained with a RELF-based features selection with a Naive Bayes classifier. Note that the RELF-based filter feature selection approach is also quite efficient on other types of HDLSS problems [26, 28].

Wrapper methods consist in selecting the subset of features that optimizes the classifier performance, typically measured on an independent validation dataset [26]. Since an exhaustive search is computationally intensive, wrapper methods usually adopt suboptimal searches, such as sequential search, or heuristics, such as genetic algorithms [25]. As far as we know, very few works have tried wrapper methods on Radiomics problems, and usually report worst results in comparison to REFL-based filter methods [29, 15]. We think it is due to the HDLSS setting that makes difficult to use an independent validation dataset for the search for the best feature subset.

Embedded methods refer to methods for with the search of the best feature subset is guided by the learning process [25]. In other words, the learning part and the feature selection part can not be seperated. Compared to wrapper methods, these feature selection methods are less computationally expensive and less prone to overfitting [30]. Several works have successfully used an embedded feature selection method, namely SVMRFE (support vector machine recursive feature elimination) [31], on Radiomics tasks. This approach differs

from the filter approaches by embedding the feature selection into the learning procedure, so that it can take the resulting classifier performance into account. In [32, 33], the authors show that SVMRFE is very accurate on Radiomics data. Note that the SVMRFE method is also known to be efficient on other kind of HDLSS problems [26].

Most of the Radiomics works use one of these feature selection methods, along with early integration (Figure 1 (a)), because they deal well with HDLSS datasets. However, they may easily filter some useful information for the classification task. The rationale behind extracting many different features is that all of them could be relevant for a given task and may complement each other for that purpose. If only a small subset of the features is chosen, certainly a lot of useful information will be lost and the heterogeneity can not be well represented.

Filter methods have the advantage not to require a validation set to perform feature selection, which is probably the reason why they are mostly used in Radiomics applications. However, they ignore the resulting classification performance and therefore, are usually less accurate than embedded and wrapper methods. On the other side, these two other families of approaches usually have a higher computational cost and are not very suitable for very low sample size problems.

2.2. Late integration methods in Radiomics

There are mainly two sorts of late integration methods: Co-training methods and Multiple Classifier Systems (MCS).

Co-training methods are semi-supervised methods that maximize the mutual agreement on two distinct views on unlabeled data by exchanging information among classifiers [34]. In practice, the original co-training algorithm may suffer in the presence of view disagreement caused by noise or view-corruption, and even very few inaccurately labeled examples can greatly affect the performance [35]. As explained in the introduction section, Radiomics datasets are usually composed of very few labeled instances and no additional unlabeled instances are available. As a consequence, co-training approaches are not straightforwardly applicable to Radiomics tasks. MCS consists in building different classifiers and combining

their outputs using techniques like majority voting. In general, the rationale behind using MCS is that combining several slightly different classifiers usually outperforms the most accurate of them [36, 37]. In multi-view learning, one classifier is usually built on each view and combined with the others afterward (Figure 1 (c)). The goal here is to exploit the complementarity of each view at the decision level. For example, in [38], the authors proposed to use five heterogeneous feature groups which represent different aspects of semantics for identifying health related messages in social media. Then, they chose five classifiers along with different combination operators to combine the predictions. Simply integrating the results of different classifiers makes MCS fast, efficient and flexible for multi-view learning.

In contrast with early integration methods, MCS-based late integration methods can exploit the complementarity of the different views by seeking to find the consensus among them. As far as we know, late integration has been applied to many multi-modalities medical problems but never on Radiomics problems so far. The reason may be that in Radiomics, most views are already HDLSS independently of one another: if MCS can deal well with multi-view data, they can however hardly deal with the HDLSS setting since they usually require a validation set for learning/optimizing the combination operator (fuser in Figure 1 (c)).

2.3. Intermediate integration methods in Radiomics

Intermediate integration methods adopt one of two principles: Subspace learning or Multi-Representation Fusion (MRF).

Subspace learning aims at finding a latent subspace shared by all the views and at fusing them together in the shared view. Subspace learning is an efficient multi-view dimensionality reduction technique, but most subspace learning methods are unsupervised and ignore the supervised information, which may lead to a subspace with weak predictive ability [39].

The main idea of MRF is to project each view into the same joint description space. By doing so, the views are directly comparable and can easily be merged together, resulting in a unique representation that is supposed to preserve the information from each view (Figure 1 (b)). Kernel-based MRF methods are the most popular for multi-view learning [40, 41].

For example, in [42], an SVM method is proposed that firstly computes separate kernels for each view and then sums the results. The multiple kernel function is given by Equation 1, where g and p stand for gene expression and phylogenetic profiles, and K is a local kernel.

$$K_{combined}(X, Y) = K(X_g, Y_g) + K(X_p, Y_p) \tag{1}$$

Methods that combine many different kernels together, linearly or not, are called MKL (Multiple Kernel Learning) methods [43]. The literature on MKL is abundant, and reviewing the different approaches is out of the scope of this paper. However, as MKL is straightforwardly applicable to multi-view learning problems [44], one of its representatives has been included in the experimental comparison of Section 5. Following the recommendation in [45], we choose the EasyMKL method since it is one of the most recent and accurate MKL methods nowadays. Let us mention that one of the approaches proposed in this paper, namely RFSVM, is quite similar to an MKL method that would use a "simple" average of kernels. As shown in [45], most of the state-of-the-art MKL methods do not necessarily significantly outperform this kind of kernel combination.

From our point of view, the MRF approach is the most suitable for Radiomics problems, since it is a multi-view learning principle, but also because it can deal well with HDLSS problems. In particular, the joint description space in which the views are projected can naturally be based on (dis)similarities, as with the kernel trick, which is powerful to overcome the high dimensionality issue, as it will be shown and discussed in the rest of this paper. Nevertheless, as far as we know, MRF has never been applied on Radiomics data, though it has already been applied to some medical pattern recognition problems.

2.4. Discussion

In a nutshell, one can see that there are mainly three methodological choices for tackling HDLSS multi-view problems: early integration with filter feature selection, MRF-based intermediate integration and MCS-based late integration. However, feature selection ignores the complementarity of the different views while MCS cannot deal well with the HDLSS prob-

lem. As for the MRF methods, they can deal with both aspects of our Radiomics context: MRF methods are naturally suited to multi-view learning, and their use of (dis)similarities is powerful to overcome the difficulty raised by the high dimensionality of data in HDLSS problems. They also make the fusion of views very straightforward since (dis)similarities (or kernels) are by construction comparable from one view to another.

However, most kernel-based MRF methods work along with SVM in the kernel space. An alternative would be to learn directly in the dissimilarity space, or in a dissimilarity-based embedded space, as explained in [46, 47]. For example, in [48], the authors use multi-modal data for Alzheimer's Disease patients. To combine four modalities together, they compute an RFD (Random Forest Dissimilarity) matrix for each modality and then fused the four matrices by averaging. Finally, multidimensional scaling is used on the joint dissimilarity matrix, and another RF (Random Forest) classifier is trained in the resulting embedded dissimilarity space. Compared to other kernel methods, RFD has the advantage of taking both feature information and class membership into consideration for computing dissimilarities between instances, thus expected to be of better quality. More details about the RFD measure is given in the next section.

3. Dissimilarity-based representation

As discussed above, intermediate integration methods can generate a better representation of data by taking advantage of the complementary information contained in each view. However, the question of how to integrate information coming from different views is a challenge because different views may have different number of features, different feature types, and are not directly comparable. Projecting each view of the data in some dissimilarity space can offer a smart solution to that issue as views become comparable (same feature type, same feature space size) and the dimension of the initial HDLSS data is reduced. In this section, we first introduce the dissimilarity-based representation, how it can be built using RF and how the RF hyperparameters influence the quality of RFD measure.

Let us recall the definition of a dissimilarity matrix. Let $\mathbf{T} = \{(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), \ldots, (\mathbf{X}_N, y_N)\}$ denotes a training set made up of N instances $\mathbf{X}_i$, each labeled with its true class y_i. De-

noting $\mathcal{X}$ the domain of the $\mathbf{X_i}$, a dissimilarity measure d is a function from $\mathcal{X}^2$ to $\mathbb{R}^+$ that estimates how dissimilar two instances are. For two given instances $\mathbf{X}_i$ and $\mathbf{X}_j$, a high value $d(\mathbf{X}_i, \mathbf{X}_j)$ means that the two instances are very "different", while on the opposite, a low $d(\mathbf{X}_i, \mathbf{X}_j)$ means they are very similar. In particular, $d(\mathbf{X}_i, \mathbf{X}_i) = 0$. For classification problems, the dissimilarity between two instances from the same class is expected to be small, while on the contrary the dissimilarity between two instances from two different classes is expected to be high.

Now, let $\mathbf{D}$ denote a $N \times N$ matrix, called a dissimilarity matrix, built from a given dissimilarity measure d and from a training set $\mathbf{T}$, and defined as in Equation (2):

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1N} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & d_{N3} & \dots & d_{NN} \end{bmatrix} \tag{2}$$

where d_{ij} denotes $d(\mathbf{X}_i, \mathbf{X}_j)$, for all $(\mathbf{X}_i, \mathbf{X}_j) \in \mathbf{T} \times \mathbf{T}$.

$\mathbf{D}$ is non-negative and respects the reflexivity condition. Such a dissimilarity matrix can be viewed as a new training set, where each training instance $\mathbf{X}_i$ is described by a vector $\{d_{i1}, d_{i2}, \dots, d_{iN}\}$. In the same way, using the dissimilarity to each of the training instances, any new instance $\mathbf{X}$ can be mapped into a N dimensional dissimilarity space DS. For HDLSS data, the dimension of this dissimilarity space is necessarily smaller than the dimension of the original feature space.

Typically, a distance measure such as the euclidean distance can be used to measure dissimilarities. However, such a measure does not capture the class membership, which is an important criterion to tell whether or not two instances are similar for the classification at hand. Compared to such a distance function without class information, class-based dissimilarity measure is more powerful, as with the Random Forest dissimilarity measure detailed in [49].

3.1. Random forest dissimilarity

Random Forest has been a very popular data mining and statistical tool for years due to its transparency and great success in classification and regression tasks as well as in unsupervised learning or active learning tasks [49, 50].

Given a training set $\mathbf{T}$, a Random Forest (RF) classifier $\mathbf{H}$ is a classifier made up of M decision trees, and is denoted as in Equation (3):

$$\mathbf{H}(\mathbf{X}) = \{h_k(\mathbf{X}), k = 1, \ldots, M\} \tag{3}$$

where $h_k(\mathbf{X})$ is a random tree grown using bagging and random feature selection. We refer the reader to [51, 52] for more details about this procedure. Note however that there exist many different RF learning methods that differ from the one in [51] by the use of different randomization techniques for growing the trees. However, we choose to use this reference method since it is the most commonly used in the literature and since each RF learned in this work is mainly used to compute the dissimilarities and not necessarily to exhibit the best accuracies.

For predicting the class of a given query instance $\mathbf{X}$ with a random tree, $\mathbf{X}$ goes down the tree structure, from its root till its terminal node. The descending path is decided by successive tests on the values of the features of $\mathbf{X}$, one per node. The prediction is given by the terminal node (or leaf node) in which $\mathbf{X}$ has landed. We refer the reader to [52] for more information about this process.

Hence if two query instances land in the same terminal node, they are likely to belong to the same class and they are also likely to share similarities in their feature vectors, since they have followed the same descending path.

The RFD measure $\mathbf{D_H}$ is inferred from an RF classifier $\mathbf{H}$, learned from $\mathbf{T}$. Let us firstly define a dissimilarity measure inferred by a decision tree $d^{(k)}$: let L_k denote the set of leaves of the k^{th} tree, and let $l_k(\mathbf{X})$ denote a function from $\mathcal{X}$ to L_k that returns the leaf node of the k^{th} tree where a given instance $\mathbf{X}$ lands when one wants to predict its class. The dissimilarity measure $d^{(k)}$, inferred by the k^{th} tree in the forest is defined as in Equation

(4): if two training instances $\mathbf{X}_i$ and $\mathbf{X}_j$ land in the same leaf of the k^{th} tree, then the dissimilarity between both instances is set to 0, else set to 1.

$$d^{(k)}(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} 0, & \text{if } l_k(\mathbf{X}_i) = l_k(\mathbf{X}_j) \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

The RFD measure $d^{(\mathbf{H})}(\mathbf{X}_i, \mathbf{X}_j)$ between $\mathbf{X}_i$ and $\mathbf{X}_j$ consists in calculating $d^{(k)}$ for each tree in the forest, and in averaging the resulting dissimilarity values over the M trees, as in Equation (5):

$$d^{(\mathbf{H})}(\mathbf{X}_i, \mathbf{X}_j) = \frac{1}{M} \sum_{k=1}^{M} d^{(k)}(\mathbf{X}_i, \mathbf{X}_j) \tag{5}$$

3.2. The parametrization of RFD

The RF learning algorithm, whether it is used for computing a dissimilarity or not, is controlled by important hyperparameters. While most of these hyperparameters have been extensively studied when RF is used as a classifier, their influence on the quality of the RFD measure is not that clear. In particular, the following hyperparameters are assumed to be crucial for having a "good" RFD measure:

- **Forest size** M: As explained in [51, 53], it is now known that the RF accuracy converges for an increasing number of trees in the forest. One can naturally wonder if the same goes for the quality of the corresponding RFD measure. As explained in the previous section, RFD is computed by averaging over the dissimilarity matrices inferred by each tree. If the number of trees is very small, say 5 trees for example, the RFD estimate would always be one of the 5 following values: 0, 0.2, 0.4, 0.6, 0.8 or 1. Obviously, this is not accurate enough for describing (dis)similarities between instances. When the number of trees increases, the RFD value is expected to be more accurate and reliable.

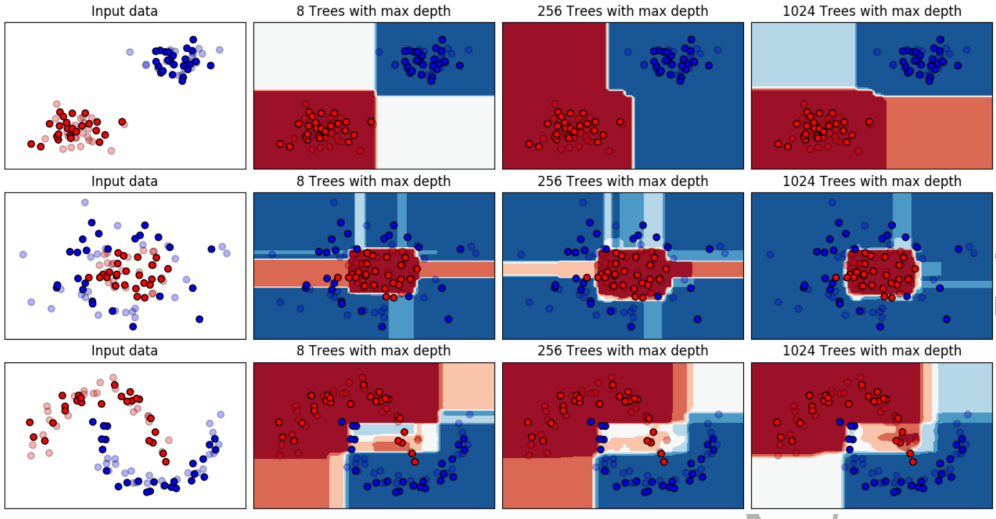
- **Tree depth** δ: The rationale behind studying the influence of the tree depth on the RFD quality is less obvious. When the node is deeper down the tree structure, it is

usually more "pure", that is to say it gathers training instances from the same class mostly. This is desirable since it means the RFD values will reflect the class membership: two instances from the same class will be considered quite similar. However, at the same time, the deeper the node, the smaller it will be, that is to say the fewer instances it will gather. As a consequence, the resulting RFD matrix is likely to be sparse, and the dissimilarity measure too loose.

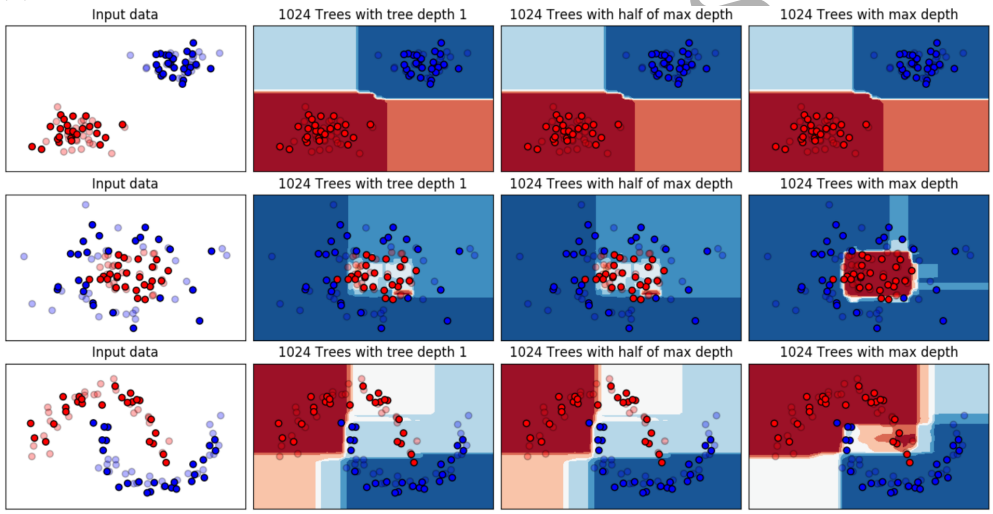
To illustrate the influence of these hyperparameters, an RF classifier is built on the dissimilarity matrix induced from different combinations of numbers of trees and tree depths on three toy datasets. The three toy datasets, composed of 100 instances, two features and two classes, have different shapes and complexities, as shown in the first column of Figures 2a and 2b: (i) the first row is a dataset with two isotropic Gaussian classes to show how the RFD measure behaves differently from a traditional Euclidean distance measure; (ii) the second row is a donut-shaped dataset, more complex with regards to similarity measures because, contrary to the RFD measure, a distance-based dissimilarity would fail to represent the class membership; and (iii) the third row is a banana-shaped dataset used to confirm that the RFD measure can properly take into account the class membership to estimate dissimilarities. The decision frontiers given by these RF are shown in the last three columns of Figure 2a and Figure 2b.

Figure 2a shows the influence of the number of trees when their depth is set to the maximum. For the dataset (i) (first row), 8 trees are enough to achieve good performance. But for the two other datasets, the influence of the number of trees can be better highlighted. For an increasing number of trees in the forest, the quality of RFD gets better as shown in Figure 2a, and it is also reflected on the decision frontier. It can be seen that for both datasets, the decision frontiers better suit to the classes (i.e. describe more and more correctly the data structure) when the number of trees increases. Therefore if we want the RFD measure to be accurate for each of the cases, it is necessary to have as many trees as possible.

Figure 2b shows the influence of the tree depth for a forest of 1024 trees. For the dataset



(a) Influence of number of trees on decision boundaries with maximum depth.



(b) Influence of tree depth on decision boundaries with 1024 trees.

Figure 2: Influence of the hyperparameters on the decision frontiers for 3 toy datasets. The transparent points are the training instances and the opaque points are the test instances. Note that in the sub-figure (b), the decision boundaries do not change in the first row because the maximum depth is 1.

(i), there is no difference in the three scatter plots because the maximum depth is equal to 1. For the two other toy datasets, it can be seen that the decision frontiers are sharper and better fit the training set when the tree is deeper. In particular, when the tree depth is not maximum, the decision boundaries are not sharp enough: this is because, in this case, the trees fail to capture the class membership of similar instances.

In summary, these results show that if we want the RFD measure to be accurate, it is

necessary to have the maximum tree depth, with a large number of trees in the forest.

3.2.1. Protocol of experiments for real-world datasets

To confirm the trends observed on the toy datasets, the hyperparameters are further studied on real-world multi-view datasets. A general description of these datasets can be found in Table 1. The first four datasets are Radiomics problems. The others 11 datasets are non-Radiomics datasets but relate to similar HDLSS multi-view applications. More details about the Radiomics datasets can be found in the work of [19]. As for the non-Radiomics datasets: LSVT is a dataset on vocal performance degradation of Parkinson's disease subjects [54]; Metabolomic contains biomarkers (CEA and TIMP), fluorescence concentration (PF) and NMR profiles for early detection of colorectal cancer [55]; BBC and BBCSport are text classification problems constructed from the news article corpora by splitting articles into related segments of text [56]; the remaining datasets (Cal7 and 20, Mfeat, NUS-WIDE2 and 3, and AWA8 and 15) are classical image classification datasets obtained using different feature extractors. Similar to the work of [57], these latter datasets have been randomly down-sampled to simulate the HDLSS setting.

All these datasets are multi-view datasets, that is to say they are supplied with several views of the same data instances. However, as the goal of this first experiment is to study the effect of hyperparameters on the quality of the RFD measure, we considered the 71 views (coming from the 15 datasets) separately, as independent datasets.

Both hyperparameters M and δ have been tested with the following values:

- Forest size $M \in \{8, 16, 32, 64, 128, 256, 512, 1024\}$: first, an RF with 1024 trees is built; the performance is then monitored with the first 8 trees, the first 16 trees, and so on, until all the 1024 trees are used in the RF. Recall that for training a random forest, trees are grown independently from each other. Therefore, retaining a subset of trees in a forest already built is just a mean to save computation time.

- Tree depth δ : an RF is firstly built with fully grown trees. For each RF, the maximum tree depth δ_{max} is computed. Then, the quality of the RFD is measured by only

Table 1: Overview of the real-world datasets used in our experiments. IR (imbalanced ratio) is the number of instances of the majority class over the number of instances of the minority class.

	#features	**#samples**	**#views**	**#classes**	**IR**
nonIDH1[19]	6746	84	5	2	3
IDHcodel[19]	6746	67	5	2	2.94
lowGrade[19]	6746	75	5	2	1.4
progression[19]	6746	84	5	2	1.68
LSVT[54]	309	126	4	2	2
Metabolomic[55]	476	94	3	2	1
Cal7[57]	3766	1474	6	7	25.74
Cal20[57]	3766	2386	6	20	24.18
Mfeat[58]	649	600	6	10	1
BBC[56]	13628	2012	2	5	1.34
BBCSport[56]	6386	544	2	5	3.16
NUS-WIDE2[59]	639	442	5	2	1.12
NUS-WIDE3[59]	639	546	5	3	1.43
AWA8[60]	10940	640	6	8	1
AWA15[60]	10940	1200	6	15	1

considering nodes above depth $i \in \{1, 2, 3, \ldots, \delta_{max}\}$, that is to say by considering that each branch of each tree has not been grown beyond depth i.

Following the conclusion of [61], the quality of the RFD measure obtained with different combinations of these hyperparameters is now assessed with a 1-Nearest Neighbor classifier (1NN) that uses the dissimilarity values instead of a distance metric. This method can well reflect the quality of a dissimilarity measure, because the idea behind 1NN is that the most similar instances should belong to the same class. A stratified random splitting strategy has been used to obtain a robust estimate of the performance of these 1NN classifiers. Each dataset has been randomly split 50 times, with 50% of the instances for training and 50% for test. A grid search has been performed on M and δ over the 50 random splits.

3.2.2. Results on real-world datasets

The results on the 71 views are presented in this section as mean and standard deviations of the classification rates over the 50 runs. To better illustrate the results, a 2D color-map is drawn for each dataset, as in Figure 3. The warm color (yellow) stands for a relatively high

quality of the RFD as measured by the 1NN accuracy while the cold color (blue) stands for relatively low quality. The y-axis corresponds to the number of trees, and the x-axis corresponds to the tree depth. For clarity concerns, only four examples are shown in Figure 3. However, these four color-maps have been chosen as they are good representatives of all the results we have obtained in this experiment.

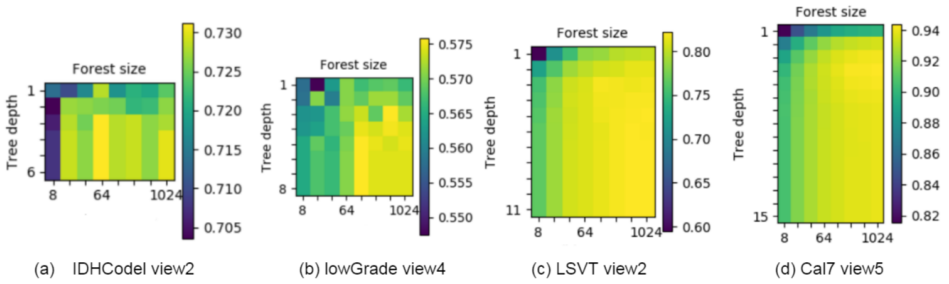


Figure 3: Color-maps of 4 of the 71 views.

Additionally, a statistical test of significance has been performed over the 71 views to state whether or not the differences observed on these color-maps are statistically significant. However, also for clarity concerns, only nine combinations of M and δ have been considered among all the possible combinations shown on the color-maps. Table 2 sums up these nine parameterization settings, that have been chosen according to the conclusions drawn on the toy datasets in the previous section. The statistical test used in this experiment is the Nemenyi post-hoc test with CD (Critical differences), as recommended in [62].

The results of the statistical test are shown in Figure 4. Over all the 71 views, the best rank is achieved by the *maxmax* setting, followed by *halfmax*, *maxhalf* and *halfhalf*. The performance of *maxmax*, *halfmax*, *maxhalf* and *halfhalf* are significantly better than the performance of *minmin*, *minhalf*, *minmax*, *halfmin* and *maxmin*, which confirms the observations made from the color-maps in Figure 3: the bottom-right corner (maximum trees, maximum depth) globally corresponds to better accuracies than the top-left corner (minimum trees, minimum depth). One can conclude that with a large number of trees, all fully grown to their maximum depth, the quality of the resulting RFD measure is guaranteed to be close to the best possible.

Table 2: The 9 combinations chosen for the statistical test

Name	Tree depth	Number of trees
minmin	1	8
minhalf	1	128
minmax	1	1024
halfmin	$\delta_{max}/2$	8
halfhalf	$\delta_{max}/2$	128
halfmax	$\delta_{max}/2$	1024
maxmin	δ_{max}	8
maxhalf	δ_{max}	128
maxmax	δ_{max}	1024

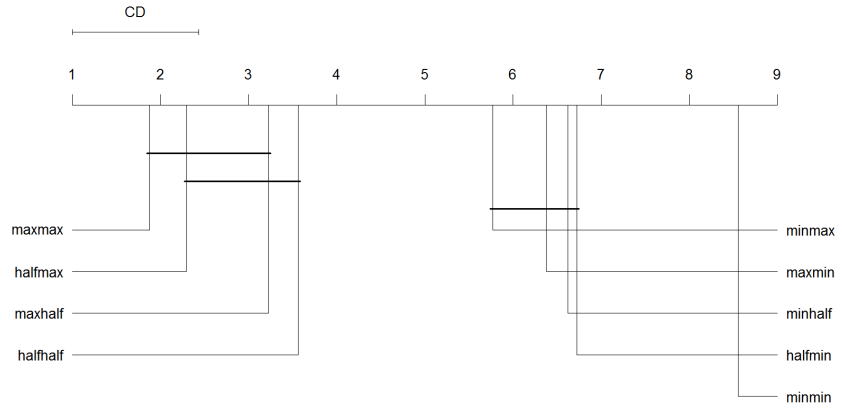


Figure 4: The post hoc test result over all views with Nemenyi test with CD (Critical value) when $\alpha = 0.05$.

Note however that one can see small differences in some of the views tested, as illustrated in Figure 3: from Figures 3 (c) and (d), it is clear that the worst RFD measure is obtained from forests with very few trees and minimum tree depth, while the best RFD measure is obtained with very large forests and maximum tree depth. In contrast, this trend is not as clear in Figures 3 (a) and (b); the worst results are not clearly located at top-left corner of the color-map, and the best results are neither located exactly at bottom-right corner. This may be due to the fact that the views on the left (Figure (a) and (b)) have a very small number of samples (67 and 75 instances respectively) and that the resulting learning phase is inherently less reliable for these cases. However, even with very few instances for learning, the trend is still observable on these figures.

3.2.3. Discussion

From this study on the parametrization, one can see that the general trend for all datasets is similar: the RFD measure is more reliable when the RF contains more trees and when these trees are fully grown. From the overall comparison on the real-world datasets, the *maxmax* setting (1024 trees with maximum depth) appeared to be better than the *maxhalf* setting (128 trees with maximum depth) but not statistically significant, which means that 128 trees already allow to obtain a quite good RFD measure for most of the views. For a better insight into this, Figure 5 shows the result of the Nemenyi post-hoc test when focusing on the number of fully grown trees. It shows that the performance gaps for forests from 256 to 1024 trees are not statistically significant. However, these differences in terms of average ranks, observable on this figure, are still important enough from our point of view to consider using more than 256 trees. Nevertheless, it is worth noting that the computational cost of learning an RF classifier is directly proportional to the number of trees [63]. Hence, in all the remaining experiments, all the RF have been learned with 512 trees as a good compromise between reliability and computational costs. And of course, all the trees have been fully grown all along these experiments.

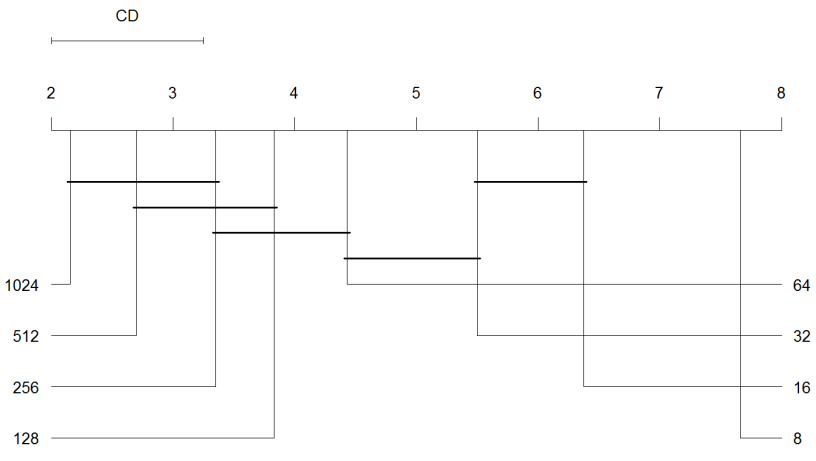


Figure 5: Comparison of RFD results with different numbers of trees and with maximum tree depth, using Nemenyi test with CD (Critical value) when $\alpha = 0.05$.

4. RFD-based multi-view learning

In the previous section, the RFD measure has been introduced and applied to each view of each dataset. For multi-view learning, one needs now to fuse the dissimilarity matrices built on each view and to learn a classifier from the resulting joint dissimilarity matrix.

A natural way to fuse the dissimilarity matrices is to compute the unweighted average matrix. For multi-view learning tasks, the training set $\mathbf{T}$ is composed of Q views: $\mathbf{T}^{(q)} = \{(\mathbf{X}_1^{(q)}, y_1), \ldots, (\mathbf{X}_N^{(q)}, y_N)\},\ q = 1..Q$. From these views, Q RFD matrices are computed following Equation 2 and noted $\{\mathbf{D}_\mathbf{H}^{(q)},\ q = 1..Q\}$. For multi-view learning, the joint dissimilarity matrix $\mathbf{D_H}$ can be computed as in Equation (6).

$$\mathbf{D_H} = \frac{1}{Q}\sum_{q=1}^{Q}\mathbf{D}_\mathbf{H}^{(q)} \tag{6}$$

According to the work in [61], learning from a dissimilarity matrix $\mathbf{D_H}$ can be done in two different ways: (i) by using the corresponding similarity matrix $\mathbf{S_H} = \mathbf{1} - \mathbf{D_H}$ as a kernel matrix in a kernel-based learning method, e.g. a SVM classifier (named RFSVM in the following and illustrated in Figure 6a) and (ii) by using the dissimilarity matrix $\mathbf{D_H}$ as a new training set (named RFDIS in the following and illustrated in Figure 6b).

Multi-view Random Forest kernel SVM (RFSVM): Instead of using traditional kernels, such as the Gaussian Radial Basis Function kernel, SVM classifiers can be efficiently trained on user-defined kernels. For example, in [64], the authors proposed a problem dependent distance measure to construct a substitution Gaussian kernel. Such a user-defined kernel can be supplied to SVM classifiers as a kernel matrix as long as it is positive semi-definite (p.s.d). For RFSVM, the joint similarity matrix $\mathbf{S_H}$ is used as a kernel matrix. The proof that it is p.s.d. is given in Appendix A. Then, given a test instance $\mathbf{X}_t$, the joint RF similarity vector $\mathbf{S}_t$, which contains the average similarities between the test instance and each training instance among different views, is given to the trained SVM for prediction.

Multi-view random forest dissimilarity (RFDIS): RFDIS consists in learning an

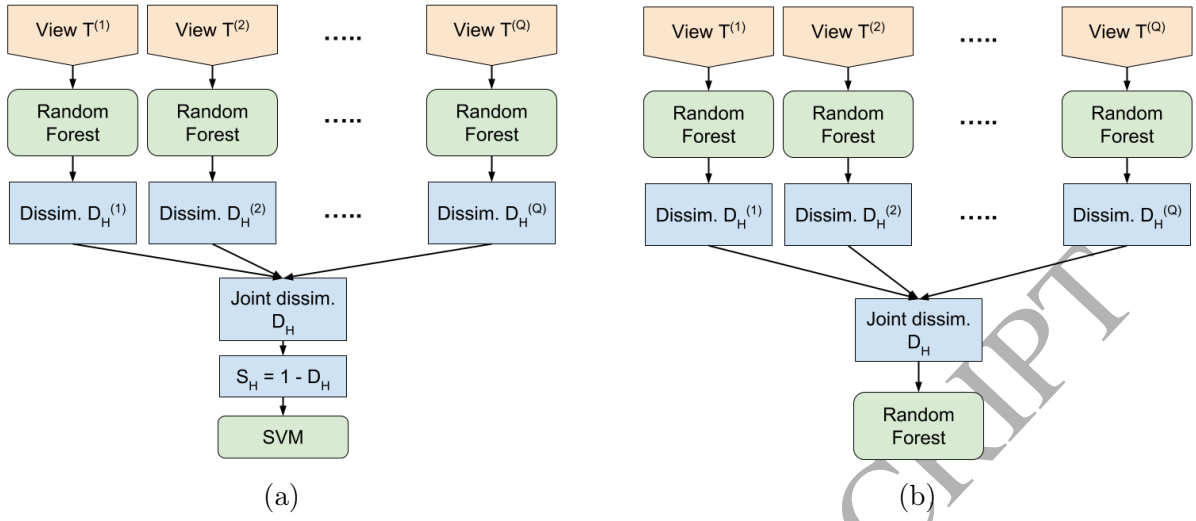


Figure 6: Flowchart of (a) the RFSVM method and (b) the RFDIS method

RF classifier $\mathbf{H}$ as if $\mathbf{D_H}$ was a new training set. The joint dissimilarity vector is seen as a feature vector, and an RF classifier is built on these new features. In other words, the original data are projected from the feature space to the dissimilarity space. In this case, the dimension of the dissimilarity space is the number of instances in the training set. For HDLSS data, it will necessarily reduce the data dimension without feature decimation. Given a test instance $\mathbf{X}_t$, the joint RF similarity vector $\mathbf{S}_t$, which contains the average similarities between the test instance and each training instance among different views, are given to the RF classifier for prediction.

As far as we know, only one method has already been proposed that perform learning from a joint RF dissimilarity matrix: the method in [48], named MDSRF in this paper. This method is similar to RFDIS except for two aspects. First, the computation of the joint similarity matrix differs in that, for RFSVM and RFDIS, it is an unweighted average combination whereas in MDSRF, it is a linear combination with weights optimized through a coarse-grained grid search. Second, RFSVM works in a kernel space and RFDIS works in a dissimilarity space, whereas MDSRF works in an embedded space obtained with a multi-dimensional scaling (MDS).

These three methods, RFSVM, RFDIS and MDSRF are compared to other state-of-the-art methods on different HDLSS multi-view learning problems in the next section.

5. Experiments and results

In this second set of experiments, the two RFD-based multi-view learning methods presented in the previous section are compared to several state-of-the-art methods, from the Radiomics and/or from the multi-view learning literature. The datasets used in this experiment are the real-world datasets used in the preliminary study (cf. Table 1) but the data are now treated as multi-view. Let us recall that all these datasets are multi-view datasets, that is to say they are made up with several views (description spaces) of the same data instances.

5.1. Experimental protocol

For this experimental validation, six methods are compared: one state-of-the-art Radiomics solution based on feature selection, namely SVMRFE [32, 33]; four intermediate integration methods, i.e. the proposed RFSVM and RFDIS presented in section 4, the MDSRF method proposed in [48] and the MKL method EasyMKL [45]; and one RFD-based late integration methods, namely LateRFDIS, from [22]. This latter method is a basic MCS architecture, which firstly builds an RFD matrix on each view, then trains an RF classifier on each of these dissimilarity matrices, and finally combines these RF classifiers by majority voting. All these methods are sum up in Table 3.

Table 3: An overview of all the methods compared in this work

Methods	Integration Method	Overview	Learning space
SVMRFE [33]	Early	Embedded feature selection; state-of-the-art in Radiomics	Reduced feature space
RFSVM	Intermediate	RFD-based MRF	Kernel space
RFDIS	Intermediate	RFD-based MRF	Dissimilarity space
MDSRF [48]	Intermediate	RFD-based MRF	Embedded dissimilarity space
EasyMKL [45]	Intermediate	Gaussian, linear and polynomial kernel based MRF	Kernel space
LateRFDIS	Late	Majority voting based MCS	Dissimilarity space

For the SVMRFE method, the number of features to select, which is a hyperparameter of the method, is set according to the total number of features following the rules described in [22]. An RF classifier is then built from the selected features. For all the RF classifiers used in this experiment, the number of trees is set to 512 as explained in section 3.2.3, while the other parameters are set by default as proposed in the *Scikit-learn* machine learning

framework [65]. As for the SVM based method, the usual hyperparameter C is used to define the penalty factor. Its value is classically set using a grid search with cross-validation. For EasyMKL, a similar grid search with cross-validation strategy is used to find the best combination of kernels among a pool of linear kernels, gaussian kernels and polynomial kernels with different hyperparameters, following the protocol given in [66].

Finally, similar to the preliminary study, a stratified random splitting procedure is repeated 10 times, with 50% of the instances for training, 50% for testing. In order to compare the methods, the mean and standard deviations of accuracy are evaluated over 10 runs.

5.2. Results and discussions

The results of this experimental comparison are given and discussed in the following, firstly for non-Radiomics datasets and secondly on the Radiomics datasets.

5.2.1. Results on non-Radiomics data

Table 4: Experimental results with 50% training data and 50% test data for non-Radiomics data

Dataset	SVMRFE	RFSVM	RFDIS	MDSRF	EasyMKL	LateRFDIS
LSVT	84.12% ±3.48	84.12% ±2.93	83.33% ±3.97	**85.07**% ±2.61	80.95% ±3.40	81.42% ±2.66
Metabolomic	63.54% ±7.53	**68.75**% ±5.10	67.71% ±5.12	67.29% ±6.94	58.33% ±5.89	64.37% ±6.55
Cal7	94.57% ±0.86	96.36% ±0.47	95.21% ±0.67	86.95% ±0.85	**96.40**% ±0.87	93.98% ±0.73
Cal20	85.06% ±1.98	88.39% ±0.34	89.12% ±0.69	64.94% ±0.52	**92.33**% ±0.59	86.15% ±0.58
Mfeat	91.13% ±4.12	**97.83**% ±0.95	97.56% ±0.99	87.90% ±2.66	95.66% ±1.54	96.56% ±1.26
BBC	74.19% ±1.76	95.63% ±0.39	92.82% ±0.67	93.37% ±1.09	**97.98**% ±0.73	88.88% ±0.50
BBCSport	78.05% ±2.88	95.56% ±0.81	81.75% ±2.70	90.69% ±2.07	**96.98**% ±0.31	81.61% ±1.96
NUS-WIDE2	91.61% ±1.26	**93.77**% ±0.85	92.49% ±2.01	92.30% ±1.50	93.15 ±0.98	91.94% ±1.28
NUS-WIDE3	76.48% ±3.07	**82.56**% ±1.78	79.41% ±1.94	80.63% ±1.34	78.33% ±1.29	78.60% ±2.26
AWA8	38.87% ±3.34	**56.90**% ±1.86	56.06% ±1.35	21.00% ±1.52	43.75% ±3.56	51.31% ±1.05
AWA15	19.31% ±1.38	37.46% ±0.78	**37.90**% ±1.49	8.2% ±0.42	24.98% ±7.90	32.35% ±1.17
Average Rank	5.04	**1.68**	2.72	4.18	3.18	4.18

The average classification rates over the 10 repetitions, along with standard deviations, are shown in Table 4. From the average ranking, it can be seen that the RFSVM method

performs globally the best among the six methods, followed by the RFDIS and EasyMKL methods, while the state-of-the-art Radiomics solution (i.e. SVMRFE) is ranked the worst.

Comparison of the multi-view solutions and the state-of-the-art Radiomics solution: From Table 4, one can see that all the multi-view methods are globally better than the feature selection method SVMRFE. To better assess the difference, a pairwise analysis based on the Sign test is computed on the number of wins, ties and losses as in [67]. The result is shown in Figure 7 (a). All the multi-view solutions are compared to SVMRFE: each vertical line indicates the critical value corresponding to a confidence level α equals to 0.10 and 0.05. If the number of wins is above these lines, the corresponding method can be considered to be significantly better than the baseline method. Figure 7 (a) shows that except for MDSRF, all the methods are significantly better than SVMRFE with $\alpha = 0.05$. RFSVM and RFDIS are the ones that win the most against SVMRFE.

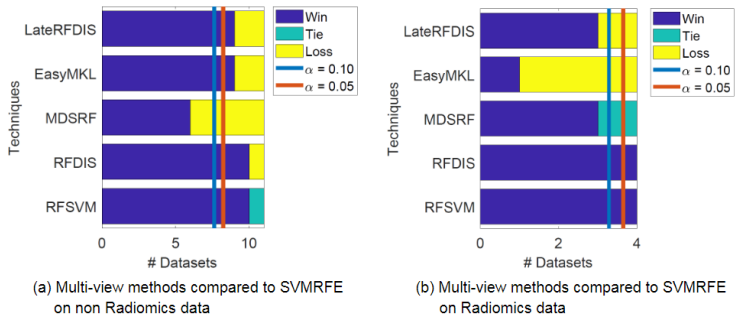


Figure 7: Pairwise comparison between multi-view solutions and feature selection methods for non-Radiomics data. The vertical lines illustrate the critical values considering a confidence level $\alpha = \{0.10, 0.05\}$.

Comparison of the RFD-based methods (RFSVM, RFDIS, MDSRF): Let us recall that each method exploits the RFD measure in the three different possible ways for dissimilarity-based pattern recognition according to [61, 68]: RFSVM uses a kernel space, RFDIS a dissimilarity space, and MDSRF an embedded space. From Table 4, it can be seen that RFSVM clearly outperforms the two other methods which indicates that kernel space seems to be the best approach. Note that the MDSRF method is quite accurate on 2-class problems, but not so much for multi-class problems, while the proposed RFSVM and RFDIS are as accurate for both. The reason may be that the MDSRF is based on an

embedded space as explained in section 3, which may suffer from a loss of information as the dimension is reduced. This information loss is more obvious for multi-class data (Cal7: 7 classes, Cal20: 20 classes, Mfeat: 10 classes, AWA8: 8 classes, AWA15: 15 classes).

Comparison of the RFSVM method and the state-of-the-art MKL method (EasyMKL): Let us recall that RFSVM and EasyMKL both adopt the same kind of kernel-based principle. From average ranking in Table 4, one can see that both RFSVM and RFDIS globally outperform EasyMKL. EasyMKL is accurate for most datasets except for the two medical datasets (LSVT and Metabolomic) with very small data size (126 samples and 94 samples respectively). This stresses that the proposed RFSVM method, as well as the RFDIS method, manage to better handle HDLSS datasets than the state-of-the-art MKL approach. Let us also recall that, contrary to EasyMKL, the RFSVM method does not require a greedy optimization of the kernel combination, neither requires to choose a priori the different kernel to use in the combination.

5.2.2. Results on Radiomics data

In the following, the previous analysis is confirmed on the real-world Radiomics datasets. The results are gathered in Table 5. By looking at the average ranking, the RFSVM method is still ranked first and the MDSRF method is ranked second. As for the feature selection method SVMRFE, it is still ranked last.

Table 5: Experimental results with 50% training data and 50% test data for Radiomics data

Dataset	SVMRFE	RFSVM	RFDIS	MDSRF	EasyMKL	LateRFDIS
nonIDH1	76.28% ±4.39	80.69% ±2.76	79.53% ±3.57	**82.55**% ±4.55	76.04% ±2.37	80.93% ±2.51
IDHcodel	73.23% ±5.50	**76.76**% ±4.52	76.47% ±3.95	73.82% ±4.26	72.35% ±2.35	76.17% ±2.06
lowGrade	62.55% ±3.36	63.95% ±4.56	63.48% ±3.76	62.55% ±5.53	64.65% ±4.26	**65.11**% ±5.20
progression	62.36% ±3.73	**65.52**% ±4.47	63.42% ±6.49	65.00% ±5.95	59.73% ±6.00	58.94% ±6.02
Average Rank	4.875	**2.000**	3.250	3.125	4.750	3.000

Comparison of the multi-view solutions and the state-of-the-art Radiomics solution: From Table 5, one can see that all the multi-view solutions are generally better

than the Radiomics solution SVMRFE. Similar to the analysis on non-Radiomics problems, a pairwise analysis based on the Sign test is also given in Figure 7 (b), and the same conclusion holds: the two proposed methods, RFSVM and RFDIS, significantly outperform SVMRFE with $\alpha = 0.05$.

Comparison of the RFD-based methods (RFSVM, RFDIS, MDSRF): Here again, the same conclusion goes with the Radiomics datasets: the RFSVM method is still the best method, followed by the MDSRF method. The MDSRF method is still slightly better on 2-class problems, which also confirms the previous conclusion that MDSRF works well for 2-class problems.

Comparison of the RFSVM method and the state-of-the-art MKL method (EasyMKL): Table 5 shows in particular that the EasyMKL method has much worse performance on Radiomics data than on non-Radiomics data, which seems to confirm that EasyMKL hardly handles very small datasets like in the medical field. The proposed RFSVM and RFDIS on the other side still work well for both Radiomics or non-Radiomics data.

5.2.3. Discussion

From the results on both non-Radiomics and Radiomics data, the following conclusions can be drawn: (i) in general, the multi-view solutions outperform the state-of-the-art Radiomics solution, and the differences are always statistically significant for the two proposed RFD-based methods; (ii) by comparing three different possibilities of learning with dissimilarity, learning in kernel space seems to be the best choice for multi-view learning problems, while one can note that the MDSRF method, that uses an embedded space, is less accurate for multi-class problems; (iii) by comparing RFSVM to MKL method, one can see that even though both methods use kernels, the RFSVM method is better than MKL, especially for very small datasets like in the Radiomics application. These results also stress that the RF kernel outperforms the traditional gaussian, linear, and polynomial kernel in the HDLSS context. Let us finally recall that the RFSVM method has the strong advantage to not require the optimization of the kernel combination, neither to choose the different kernels to use beforehand.

6. Conclusion

In this work, by treating the Radiomics application as an HDLSS multi-view learning problem, it has been shown that the classical feature selection approach adopted by most works in the Radiomics literature is not the most appropriate learning principle for inferring an accurate predictive model. By comparing such a state-of-the-art Radiomics method, namely SVMRFE, to multi-view learning methods, we have shown the potential of the latter, in particular for the intermediate integration category of multi-view methods.

The key idea of intermediate integration methods is to fuse the different views on a feature level, before learning. To do so, a dissimilarity-based representation has been proposed to allow for the views to be projected in the same description space and to be fused straightforwardly. This dissimilarity space is obtained with the Random Forest Dissimilarity measure that captures the similarities between instances, from their initial representation in each view as well as from their class membership. A preliminary experiment has been proposed to better understand how the RFD measure behaves according to the most important hyperparameters. We have shown that when there are more trees in the forest, and the trees are deeper, the resulting RFD estimate will be more accurate.

In the second set of experiments, five multi-view methods (four intermediate integration methods and one late integration method) have been compared to the Radiomics state-of-the-art method, on several HDLSS multi-view datasets, including four Radiomics datasets. The results have shown that the multi-view solutions are globally better, but only the two proposed intermediate integration methods, namely RFSVM and RFDIS, significantly outperform the state-of-the-art Radiomics solution SVMRFE. These proposed approaches, that use two well-known principles of dissimilarity-based pattern recognition (kernel and dissimilarity spaces), have also been compared to a third RFD-based method, MDSRF, that uses a third dissimilarity-based representation (embedded space), and the result shows that the kernel space is globally the best option in the HDLSS multi-view setting. Finally, the comparison has been extended with a Multiple Kernel Learning method, namely EasyMKL, known to be efficient and straightforward for application to multi-view learning. The results

show that the RFSVM method is more accurate than EasyMKL while avoiding a greedy optimization for the combination of a pool of different predefined kernels.

In this work, the RFD measure used for each view shares the same parameter settings, while we think it could be further fruitful to make the RFD measure suit to the specificities of each view. One future work will thus be focused on how to better set the parameters for the RFD measure in a more dynamic fashion. Secondly, the two dissimilarity-based intermediate integration methods treat all the views with the same importance, while a weighted combination could also be used to generate a better joint dissimilarity matrix. Finally, all the datasets we tested in this work are complete, without missing values nor missing views, while it is a preponderant issue for multi-view problems, especially in the Radiomics field. How to deal with this kind of problem is also one of the tasks in our future works.

Acknowledgment

This work is part of the DAISI project, co-financed by the European Union with the European Regional Development Fund (ERDF) and by the Normandy Region.

References

References

[1] E. Florez, A. Fatemi1, P. P. Claudio, C. M. Howard, Emergence of radiomics: Novel methodology identifying imaging biomarkers of disease in diagnosis, response, and progression, SM Journal of Clinical and Medical Imaging 4 (1) (2018) 1019.

[2] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, et al., Radiomics: the process and the challenges, Magnetic Resonance Imaging 30 (9) (2012) 1234–1248.

[3] G. Lee, H. Y. Lee, H. Park, M. L. Schiebler, E. J. van Beek, Y. Ohno, J. B. Seo, A. Leung, Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art, European Journal of Radiology 86 (2017) 297–307.

[4] S. E. Viswanath, P. Tiwari, G. Lee, A. Madabhushi, Dimensionality reduction-based fusion approaches for imaging and non-imaging biomedical data: concepts, workflow, and use-cases, BMC Medical Imaging 17 (1) (2017) 2.

[5] H. Aerts, E. R. Velazquez, R. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, et al., Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nature Communications 5 (4006) (2014) 1–8.

[6] C. Parmar, P. Grossmann, D. Rietveld, M. M. Rietbergen, P. Lambin, H. J. Aerts, Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer, Frontiers in Oncology 5 (2015) 272.

[7] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, H. J. Aerts, Machine learning methods for quantitative radiomic biomarkers, Scientific Reports 5 (2015) 13087.

[8] S. H. Hawkins, J. N. Korecki, Y. Balagurunathan, Y. Gu, V. Kumar, S. Basu, L. O. Hall, D. B. Goldgof, R. A. Gatenby, R. J. Gillies, Predicting outcomes of nonsmall cell lung cancer using CT image features, IEEE Access 2 (2014) 1418–1426.

[9] W. Wu, C. Parmar, P. Grossmann, J. Quackenbush, P. Lambin, J. Bussink, R. Mak, H. J. Aerts, Exploratory study to identify radiomics classifiers for lung cancer histology, Frontiers in Oncology 6 (2016) 71.

[10] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, et al., Radiomics: extracting more information from medical images using advanced feature analysis, European Journal of Cancer 48 (4) (2012) 441–446.

[11] X. Fave, D. Mackin, J. Lee, J. Yang, L. Zhang, et al., Computational resources for radiomics, Translational Cancer Research 5 (4) (2016) 340–348.

[12] M. Scrivener, E. E. de Jong, J. E. van Timmeren, T. Pieters, B. Ghaye, X. Geets, Translational Cancer Research 5 (4) (2016) 398–409.

[13] T. P. Coroller, V. Agrawal, V. Narayan, Y. Hou, P. Grossmann, S. W. Lee, R. H. Mak, H. J. Aerts, Radiomic phenotype features predict pathological response in non-small cell lung cancer, Radiotherapy and Oncology 119 (3) (2016) 480–486.

[14] G. Carneiro, L. Oakden-Rayner, A. P. Bradley, J. Nascimento, L. Palmer, Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography, in: 14th International Symposium on Biomedical Imaging (ISBI), 2017, pp. 130–134.

[15] H. Farhidzadeh, J. Y. Kim, J. G. Scott, D. B. Goldgof, L. O. Hall, L. B. Harrison, Classification of progression free survival with nasopharyngeal carcinoma tumors, in: SPIE Medical Imaging, International Society for Optics and Photonics, 2016, pp. 97851I–97851I.

[16] A. Cameron, F. Khalvati, M. A. Haider, A. Wong, Maps: a quantitative radiomics approach for prostate

cancer detection, IEEE Transactions on Biomedical Engineering 63 (6) (2016) 1145–1156.

[17] Y. Balagurunathan, Y. Gu, H. Wang, V. Kumar, O. Grove, S. Hawkins, J. Kim, D. B. Goldgof, L. O. Hall, R. A. Gatenby, et al., Reproducibility and prognosis of quantitative features extracted from CT images, Translational Oncology 7 (1) (2014) 72–87.

[18] T. P. Coroller, P. Grossmann, Y. Hou, E. R. Velazquez, R. T. Leijenaar, G. Hermann, P. Lambin, B. Haibe-Kains, R. H. Mak, H. J. Aerts, CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma, Radiotherapy and Oncology 114 (3) (2015) 345–350.

[19] H. Zhou, M. Vallières, H. X. Bai, C. Su, H. Tang, D. Oldridge, Z. Zhang, B. Xiao, W. Liao, Y. Tao, et al., MRI features predict survival and molecular markers in diffuse lower-grade gliomas, Neuro-Oncology 19 (6) (2017) 862–870.

[20] S. Leger, A. Zwanenburg, K. Pilz, F. Lohaus, A. Linge, K. ZÃűphel, J. Kotzerke, A. Schreiber, I. Tinhofer, V. Budach, A. Sak, M. Stuschke, P. Balermpas, C. RÃűdel, U. Ganswindt, C. Belka, S. Pigorsch, S. E. Combs, D. MÃűnnich, D. Zips, M. Krause, M. Baumann, E. G. C. Troost, S. LÃűck, C. Richter, A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling, Nature Research Scientific Reports 7.

[21] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, arXiv preprint arXiv:1304.5634.

[22] H. Cao, S. Bernard, L. Heutte, R. Sabourin, Dissimilarity-based representation for radiomics applications, First International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI).

[23] A. Serra, M. Fratello, V. Fortino, G. Raiconi, R. Tagliaferri, D. Greco, Mvda: a multi-view genomic data integration methodology, BMC Bioinformatics 16 (1) (2015) 261.

[24] T. Li, S. Zhu, Q. Li, M. Ogihara, Gene functional classification by semi-supervised learning from heterogeneous data, in: ACM Symposium on Applied Computing, ACM, 2003, pp. 78–82.

[25] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Computers & Electrical Engineering 40 (1) (2014) 16–28.

[26] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, A review of feature selection methods on synthetic data, Knowledge and Information Systems 34 (3) (2013) 483–519.

[27] A. R. S. Parmezan, H. D. Lee, F. C. Wu, Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework, Expert Systems with Applications 75 (2017) 1–24.

[28] Z. Zhao, L. Wang, H. Liu, J. Ye, On similarity preserving feature selection, IEEE Transactions on Knowledge and Data Engineering 25 (3) (2013) 619–632.

[29] S. Basu, L. Hall, D. Goldgof, Y. Gu, V. Kumar, J. Choi, Developing a classifier model for lung tumors in ct-scan images, in: IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2011, pp. 1306–1318.

[30] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinfor-

matics 23 (19) (2007) 2507–2517.

[31] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning 46 (1-3) (2002) 389–422.

[32] J. Wang, C.-J. Wu, M.-L. Bao, J. Zhang, X.-N. Wang, Y.-D. Zhang, Machine learning-based analysis of mr radiomics can help to improve the diagnostic performance of pi-rads v2 in clinically relevant prostate cancer, European Radiology 27 (10) (2017) 4082–4090.

[33] X. Zhang, X. Xu, Q. Tian, B. Li, Y. Wu, Z. Yang, Z. Liang, Y. Liu, G. Cui, H. Lu, Radiomics assessment of bladder cancer grade using texture features from diffusion-weighted imaging, Journal of Magnetic Resonance Imaging 46 (5) (2017) 1281–1288.

[34] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: Recent progress and new challenges, Information Fusion 38 (2017) 43–54.

[35] S. Sun, A survey of multi-view machine learning, Neural Computing and Applications 23 (7-8) (2013) 2031–2038.

[36] L. I. Kuncheva, J. C. Bezdek, R. P. Duin, Decision templates for multiple classifier fusion: an experimental comparison, Pattern Recognition 34 (2) (2001) 299–314.

[37] D. M. Tax, M. Van Breukelen, R. P. Duin, J. Kittler, Combining multiple classifiers by averaging or by multiplying?, Pattern Recognition 33 (9) (2000) 1475–1485.

[38] S. Tuarob, C. S. Tucker, M. Salathe, N. Ram, An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages, Journal of Biomedical Informatics 49 (2014) 255–268.

[39] N. Chen, J. Zhu, E. P. Xing, Predictive subspace learning for multi-view data: a large margin approach, in: Advances in Neural Information Processing Systems (NIPS), 2010, pp. 361–369.

[40] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, W. S. Noble, Kernel-based data fusion and its application to protein function prediction in yeast, in: Biocomputing, World Scientific, 2003, pp. 300–311.

[41] P. Tiwari, J. Kurhanewicz, A. Madabhushi, Multi-kernel graph embedding for detection, gleason grading of prostate cancer via mri/mrs, Medical Image Analysis 17 (2) (2013) 219–235.

[42] P. Pavlidis, J. Weston, J. Cai, W. N. Grundy, Gene functional classification from heterogeneous data, in: Fifth Annual International Conference on Computational Biology (ICCB), ACM, 2001, pp. 249–255.

[43] M. Gönen, E. Alpaydın, Multiple kernel learning algorithms, Journal of Machine Learning Research 12 (Jul) (2011) 2211–2268.

[44] V. Cheplygina, D. M. Tax, M. Loog, Multiple instance learning with bag dissimilarities, Pattern Recognition 48 (1) (2015) 264–275.

[45] F. Aiolli, M. Donini, Easymkl: a scalable multiple kernel learning algorithm, Neurocomputing 169

(2015) 215–224.

[46] E. Pękalska, R. P. Duin, P. Paclík, Prototype selection for dissimilarity-based classifiers, Pattern Recognition 39 (2) (2006) 189–208.

[47] E. Pękalska, R. P. Duin, Dissimilarity representations allow for building good classifiers, Pattern Recognition Letters 23 (8) (2002) 943–956.

[48] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, D. Rueckert, A. D. N. Initiative, et al., Random forest-based similarity measures for multi-modal classification of alzheimer's disease, NeuroImage 65 (2013) 167–175.

[49] T. Shi, S. Horvath, Unsupervised learning with random forest predictors, Journal of Computational and Graphical Statistics 15 (1) (2006) 118–138.

[50] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems, Journal of Machine Learning Research 15 (1) (2014) 3133–3181.

[51] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[52] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2) (2016) 197–227.

[53] S. Bernard, S. Adam, L. Heutte, Using random forests for handwritten digit recognition, in: Ninth International Conference on Document Analysis and Recognition (ICDAR), Vol. 2, IEEE, 2007, pp. 1043–1047.

[54] A. Tsanas, M. A. Little, C. Fox, L. O. Ramig, Objective automatic assessment of rehabilitative speech treatment in parkinson's disease, IEEE Transactions on Neural Systems and Rehabilitation Engineering 22 (1) (2014) 181–190.

[55] R. Bro, H. J. Nielsen, F. Savorani, K. Kjeldahl, I. J. Christensen, N. Brünner, A. J. Lawaetz, Data fusion in metabolomic cancer diagnostics, Metabolomics 9 (1) (2013) 3–8.

[56] R. Xia, Y. Pan, L. Du, J. Yin, Robust multi-view spectral clustering via low-rank and sparse decomposition., in: 28th Conference on Artificial Intelligence, 2014, pp. 2149–2155.

[57] Y. Li, F. Nie, H. Huang, J. Huang, Large-scale multi-view spectral clustering via bipartite graph, in: 29th AAAI Conference on Artificial Intelligence, 2015, pp. 2750–2756.

[58] A. Frank, A. Asuncion, et al., UCI machine learning repository (2010).

[59] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y.-T. Zheng, Nus-wide: A real-world web image database from national university of singapore, in: ACM Conference on Image and Video Retrieval (CIVR), Santorini, Greece., 2009.

[60] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 951–958.

[61] R. P. Duin, E. Pekalska, The dissimilarity space: Bridging structural and statistical pattern recognition,

Pattern Recognition Letters 33 (7) (2012) 826–832.

[62] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (Jan) (2006) 1–30.

[63] G. Louppe, Understanding random forests: From theory to practice, Ph.D. thesis, University of Liège, Belgium (2014).

[64] B. Haasdonk, C. Bahlmann, Learning with distance substitution kernels, in: Pattern Recognition - Proceedings of the 26th DAGM Symposium, Springer, 2004, pp. 220–227.

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[66] A. Rakotomamonjy, F. R. Bach, S. Canu, Y. Grandvalet, Simplemkl, Journal of Machine Learning Research 9 (Nov) (2008) 2491–2521.

[67] R. M. Cruz, D. V. Oliveira, G. D. Cavalcanti, R. Sabourin, Fire-des++: Enhanced online pruning of base classifiers for dynamic ensemble selection, Pattern Recognition 85 (2019) 149–160.

[68] E. Pekalska, R. P. Duin, Classifiers for dissimilarity-based pattern recognition, in: 15th International Conference on Pattern Recognition (ICPR), Vol. 2, IEEE, 2000, pp. 12–16.

[69] E. Pekalska, P. Paclik, R. P. Duin, A generalized kernel approach to dissimilarity-based classification, Journal of Machine Learning Research 2 (Dec) (2001) 175–211.

[70] A. N. Letchford, M. M. Sørensen, Binary positive semidefinite matrices and associated integer polytopes, Mathematical Programming 131 (1) (2012) 253–271.

Appendix A. Proof that the joint similarity matrix used in the RFSVM method is positive semi-definite

In the following, a proof that the RF similarity matrix is symmetric and positive semi-definite (p.s.d) is detailed. These two properties are essential since they ensure that such a matrix can be used as a kernel matrix in kernel methods like non-linear SVM [69]. Note that RF similarity matrices are easily obtained from the RFD matrices defined in the previous section, by $\mathbf{S_H} = \mathbf{1} - \mathbf{D_H}$.

Let us firstly recall the two following theorems from [70]:

Theorem Appendix A.1. *If both $\boldsymbol{A}$ and $\boldsymbol{B}$ are two p.s.d. matrices then so is $\boldsymbol{A} + \boldsymbol{B}$. This follows immediately from the equation $x^T(\boldsymbol{A}+\boldsymbol{B})x = x^T\boldsymbol{A}x + x^T\boldsymbol{B}x \geq 0$. Consequently any sum of p.s.d. matrices is p.s.d.*

Theorem Appendix A.2. *A symmetric binary matrix* $\boldsymbol{MA} \in (0,1)^{n\times n}$*, with* $n \geq 3$*, is p.s.d. if and only if it satisfies the following inequalities:*

$$\boldsymbol{MA}_{ij} \leq \boldsymbol{MA}_{ii}, \qquad (1 \leq i < j \leq n) \tag{A.1}$$

$$\boldsymbol{MA}_{il} + \boldsymbol{MA}_{jl} \leq \boldsymbol{MA}_{ll} + \boldsymbol{MA}_{ij}, \qquad (1 \leq i < j \leq n, l \neq i, j) \tag{A.2}$$

Using these theorems, let us demonstrate that the similarity matrix inferred by a single tree k, noted $\mathbf{S}^{(k)}$ is p.s.d. From Equation 4, one can see that $\mathbf{S}^{(k)}$ has the following properties:

- $\mathbf{S}^{(k)}$ is a symmetric matrix with principal diagonal values equal to 1.
- The off diagonal entries in $\mathbf{S}^{(k)}$ are either 0 or 1.

One can reasonably consider that the number of training instances available is greater than 3, and as a consequence, that $\mathbf{S}^{(k)}$ is a symmetric binary matrix $\in (0,1)^{n\times n}$, with $n \geq 3$. According to Theorem Appendix A.2, for this matrix to be p.s.d., it needs to satisfy both Equations (A.1) and Equation (A.2):

1. As $\mathbf{S}^{(k)}$ is a symmetric binary matrix with principal diagonal values $\mathbf{S}^{(k)}_{ii}$ equal to 1, hence $\mathbf{S}^{(k)}_{ij} \leq \mathbf{S}^{(k)}_{ii}$, which satisfies Equation (A.1).
2. To prove $\mathbf{S}^{(k)}$ satisfies Equation (A.2), two situations need to be considered:
   (a) If $\mathbf{S}^{(k)}_{ij} = 1$, then $\mathbf{S}^{(k)}_{ll} + \mathbf{S}^{(k)}_{ij} = 2$. Since $\mathbf{S}^{(k)}_{il} \leq 1$ and $\mathbf{S}^{(k)}_{lj} \leq 1$, then $\mathbf{S}^{(k)}_{il} + \mathbf{S}^{(k)}_{lj} \leq \mathbf{S}^{(k)}_{ll} + \mathbf{S}^{(k)}_{ij}$.
   (b) If $\mathbf{S}^{(k)}_{ij} = 0$, then $\mathbf{S}^{(k)}_{ll} + \mathbf{S}^{(k)}_{ij} = 1$. At the same time, $\mathbf{S}^{(k)}_{ij} = 0$ means that the i^{th} and j^{th} instances fall in different terminal nodes, which implies that $\mathbf{S}^{(k)}_{il}$ and $\mathbf{S}^{(k)}_{lj}$ can not be both equal to 1. Thus $\mathbf{S}^{(k)}_{il} + \mathbf{S}^{(k)}_{lj}$ is necessarily less or equal to 1 and as a consequence, $\mathbf{S}^{(k)}$ also satisfies Equation (A.2).

This proves that $\mathbf{S}^{(k)}$ meets the requirements of Theorem Appendix A.2, and is a p.s.d. matrix. It follows from Theorem Appendix A.1 that the sum of all $\mathbf{S}^{(k)}, \forall k = 1..M$, is also p.s.d., meaning the RF similarity matrix $\mathbf{S_H}$ or any linear combination of $\mathbf{S_H}$ is also p.s.d.

Biography

Hongliu CAO received his double M.S degrees in Telecommunication from NJUST China and in Information systems from Mines Nancy France. He is currently a Phd student in double diplomas program between LITIS, France and LIVIA, ETS, Canada. His research interests are on machine learning, data analysis, Radiomics, ensemble methods and dissimilarity. His Phd focuses on using machine learning methods in the treatment of cancer.

Simon BERNARD received his Ph.D. degree in Computer Science from the University of Rouen, France, in 2009. He currently works as an Associate Professor at the University of Rouen and as a member of the Machine Learning research team of the LITIS laboratory. His research work concerns Machine Learning, and in particular Ensemble Learning, with a focus on classification and weak supervised learning.

Robert Sabourin joined the staff of the École de technologie supérieure, Université du Québec 1983, in Montréal where he co-founded the Dept. of Automated Manufacturing Engineering where he is currently full Professor. His research interests are in the areas of adaptive biometric systems, adaptive surveillance systems in dynamic environments, intelligent watermarking systems, evolutionary computation and ensemble learning.

Laurent Heutte is a Full Professor in Computer Engineering at the University of Rouen, France. He is a member of LITIS lab and NormaSTIC (FR CNRS 3638). His research interests cover all areas of pattern recognition and machine learning, with a focus on statistical techniques, classifier ensembles and random forests.