A Multimodal Feature Selection Method for Remote Sensing Data Analysis Based on Double Graph Laplacian Diagonalization

Eduard Khachatrian, *Student Member, IEEE*, Saloua Chlaily, *Member, IEEE*, Torbjørn Eltoft, *Member, IEEE*, and Andrea Marinoni, *Senior Member, IEEE*

***Abstract*—When dealing with multivariate remotely sensed records collected by multiple sensors, an accurate selection of information at the data, feature, or decision level is instrumental in improving the scenes' characterization. This will also enhance the system's efficiency and provide more details on modeling the physical phenomena occurring on the Earth's surface. In this article, we introduce a flexible and efficient method based on graph Laplacians for information selection at different levels of data fusion. The proposed approach combines data structure and information content to address the limitations of existing graph-Laplacian-based methods in dealing with heterogeneous datasets. Moreover, it adapts the selection to each homogenous area of the considered images according to their underlying properties. Experimental tests carried out on several multivariate remote sensing datasets show the consistency of the proposed approach.**

***Index Terms*—Gaussian kernel (GK), graph Laplacians, multimodal remote sensing, mutual information (MI), unsupervised information selection.**

I. Introduction

OVER the past several decades, satellite imagery has become a crucial source in providing a vast amount of information about the Earth's surface. Thanks to technological advances, a region of interest (ROI) can be monitored by various sensors characterized by different acquisition techniques (*modalities*), using different spectral, temporal, or spatial resolutions [1]. The information provided by multiple sensors grasps different aspects of the area of interest. For instance, hyperspectral images might reveal the material content of the observed region, while synthetic aperture radar (SAR) complements the capabilities of optical imaging by reporting the topographic (interferometry) and surface roughness information, and the light detection and ranging (LiDAR) technology provides highly accurate measurements of the vertical height of structures. Accordingly, robust characterization of the Earth's surface can be achieved by combining data coming from different modalities to obtain useful insight into various aspects of the underlying surface [2].

The combination of multimodal datasets raises several challenges [1], [2]. These challenges are limited not only to dealing with the heterogeneity of the multimodal images in terms of temporal, spatial, and radiometric resolutions, sizes, and data types [2], but also to accurately selecting the relevant information that maximizes the benefits of the multimodal analysis. By expanding the size of a dataset, we are simultaneously increasing the complexity of the records to be analyzed, especially when it is multimodal. Hence, the considered algorithms might fail to capture the data's underlying structure, i.e., not achieving an accurate and robust characterization of the physical phenomena occurring on Earth's surface. Indeed, it has been shown that increasing the number of modalities without properly addressing an investigation of the significance and reliability of the data may deteriorate the analysis [3], [4]. This may, therefore, represent a strong limiting factor to the use of multimodal remote sensing data analysis in practical scenarios, as well as to its actual impact in operational frameworks within private and public sectors [1].

In fact, not all information provided by several sensors is valuable; it can be redundant, corrupted, or unnecessary for the given task [1], [2]. Therefore, to get the most use of a multimodal dataset, it is crucial to select only relevant information. In this way, it is expected that the accuracy of the analysis will increase, whereas the time complexity will be reduced. Consequently, to improve the knowledge about an observed area, there is a need to develop an automatic method to select the relevant information from various sensors [1]–[3].

Dimensionality reduction has been proven as an effective tool to tackle these issues in remote sensing data analysis [2], [3], [5]–[7]. Feature extraction and feature selection methods are able to strongly enhance the quality of understanding and assessment of physical–chemical phenomena on the ground, especially when data collected by means of homogeneous acquisition techniques (i.e., sensors with similar properties of the sensing devices) are analyzed. Nonetheless, traditional methods

Manuscript received March 8, 2021; revised May 28, 2021 and September 17, 2021; accepted October 27, 2021. Date of publication November 2, 2021; date of current version November 22, 2021. This work was supported in part by the Centre for Integrated Remote Sensing and Forecasting for Arctic Operations (CIRFA), Research Council of Norway, under Grant 237906 and in part by the Automatic Multi-sensor Remote Sensing for Sea Ice Characterization (AMUSIC) "Polhavet" flagship project 2020. *(Corresponding author: Eduard Khachatrian.)*

The authors are with the Department of Physics and Technology, University of Tromsø—The Arctic University of Norway, N-9037 Tromsø, Norway (e-mail: eduard.khachatrian@uit.no; saloua.chlaily@uit.no; torbjorn.eltoft@uit.no; andrea.marinoni@uit.no).

Digital Object Identifier 10.1109/JSTARS.2021.3124308

for dimensionality reduction might fail in capturing the details of elements, materials, and dynamic phenomena on Earth's surface when multimodal datasets are explored [2], [3].

We note that the term *feature* is commonly used in different fields such as classification methodologies, pattern recognition, and texture analysis. However, in our work, to prevent confusion with textural features, we introduce the notation *attribute* from information theory, which refers to directly measured quantities as, e.g., optical/hyperspectral/LiDAR reflectance across the electromagnetic spectrum, and additional parameters such as textural features.

In the case of multimodal datasets, which reside on a nonlinear manifold, graphs are the appropriate representation of the data. The graph is composed of the multimodal attributes as nodes, and their similarities will give the weights to their connecting edges. The dimensionality reduction is carried out by determining similar nodes and picking a representative attribute from each group. The graph partition reveals the pattern of the attributes; as such, the chosen attributes preserve the structure of the graph. The problem of graph partitioning or clustering to group similar nodes is nondeterministic polynomial-time hard (NP-hard), but it can be approximated via several techniques, such as spectral clustering (SC) [8]. In fact, the graph structure can be understood and analyzed via the Laplacian of the adjacency matrix that summarizes the nodes' similarities. In particular, the eigenvectors of the Laplacian matrix, associated with the lowest eigenvalues, reveal the structure information of the graph [9]. However, in the case of graphs of heterogeneous degrees, where the nodes interact differently, the graph's eigenvalues scatter across the spectrum. Accordingly, it will be hard to distinguish the lowest eigenvalues and determine the informative eigenvectors, which will undermine the attributes selection's pertinence and efficacy [10]–[12].

In this article, we introduce an approach to information selection in multimodal remote sensing datasets that relies on a representation based on graph Laplacians. While the existing works using graph Laplacians exploit the attributes' structure using kernels as similarity measures, we additionally consider the attribute's information content. As such, we jointly exploit mutual information (MI) and the Gaussian kernel (GK) similarity metrics to capture the most relevant attributes within the records. The two similarity measures are applied at different detail levels. The MI is used globally, considering all the pixels within the images to ensure a better estimation of the attributes' shared information. On the other hand, the GK is employed locally to preserve the particularity of homogeneous areas within the images. Accordingly, different attributes are selected for different parts of an image that might belong to different classes or be measured under different conditions (i.e., different noise levels, clouds coverage, etc.).

The main motivation of this work is, thus, the limitation of classic graph-Laplacian-based approaches at separating the attributes when they are heterogeneous, as it has been shown in [10]–[12]. Nevertheless, the joint employment of the MI and the GK at different scales ensures a better separability of the attributes and, hence, a more precise selection. Accordingly, the proposed approach guarantees high accuracy of the analysis and reduces the computational complexity so that the potential of multimodal remote sensing data analysis can be exploited in multiple applications.

The rest of this article is organized as follows. Section II reports a brief summary of the main methods for information selection in remote sensing data analysis and the main contribution of the proposed approach. Section III provides details of the proposed architecture. Section IV presents an experimental validation of the proposed method. Finally, Section V concludes this article.

For notational convenience, random scalars are denoted by lowercase letters, e.g., z. Random vectors are designated by bold lowercase letters, e.g., $\mathbf{z}$. Bold uppercase letters refer to matrices, e.g., $\mathbf{A}$. $|\mathbf{A}|$ and $\mathrm{Tr}(\mathbf{A})$ denote the determinant and trace of the matrix $\mathbf{A}$, respectively. $\mathrm{diag}\{d_1, \ldots, d_N\}$ refers to a diagonal matrix whose diagonal elements are $d_1, \ldots, d_N$ starting from the upper left. The $\mathrm{ddiag}(\mathbf{A})$ operator is set to zero the off-diagonal entries of $\mathbf{A}$.

II. Background and Motivation

A. Existing Work

In order to select the most informative subset of attributes and discard the irrelevant ones, it is possible to use several dimensionality reduction methods. Generally, dimensionality reduction methods can be separated into two main approaches: attribute extraction and attribute selection [13], [14].

1) *Attribute extraction* reduces the dimensionality by projecting the original data into a lower dimensional space [15], [16]. As such, the separability of the data is increased but at the expense of physical interpretability, which is essential in remote sensing analysis. Among the methods of attribute extraction, we may cite, for instance, principal component analysis (PCA) [17] and decision boundary feature extraction (DBFE) [5]. PCA converts a set of attributes of potentially correlated variables into a set of linearly uncorrelated variables, called principal components. It projects the original set into a lower dimensional space spanned by the principal eigenvectors of data's covariance matrix. Thus, it reduces the size of the original set while preserving its variance [17]. DBFE is a supervised approach that uses the training set to determine the decision boundary between classes. The eigenvectors of the decision boundary matrix determine the direction of projection of the original set of attributes. As such, it provides a minimum number of transformed attributes that achieve the same accuracy as the original set [5].
2) *Attribute selection* reduces dimensionality by selecting the most informative subset of records preserving the characteristics of the original data without working on a different space [18]. Attribute selection determines a subset of the original set that is more relevant according to some criteria, such as information, similarity, or correlation. The methods for attribute selection can be divided into three categories: ranking, searching, and clustering.

a) *Ranking methods* sort the attributes with respect to a given criterion and select the most significant elements [6], [19]. They are very efficient, but they might not be very precise because they do not consider the relationships among the attributes. Among this family of attribute selection methods, we can cite Fisher score for attribute selection (FIS). FIS is a supervised approach that selects the subset of attributes with a large Fisher score. The Fisher score measures the ability of each attribute to reduce the intraclass distance while increasing the interclass distance [20].

b) *Searching methods* select the optimal subset in an incremental, removal, or update manner using a search method, such as a genetic algorithm (GA) [21] or branch and bound [22]. This class of attribute selection algorithms is more accurate than ranking methods since it considers the interaction between the data. However, such methods are limited by the size of the searching space. In the case of large datasets, computation time significantly increases, and the searching methods fail to achieve optimal results. Here, we can highlight forward attribute selection (FS) [23], orthogonal branch and bound (OBB) [22], and GA [21]. FS determines the optimal subset in an incremental fashion. The algorithm starts with a minimum number of attributes, and with each new step, it adds one attribute that improves the accuracy until no further improvement is noticed [23]. The OBB is a backtracking attribute selection algorithm. It is based on the assumption that the adopted criterion function fulfills the monotonicity condition. Hence, it guarantees to find the optimal subset while omitting many attribute subset evaluations. The branching step consists of constructing the tree such that the subtree of each level is constructed by deleting one attribute until the required number of attributes is reached. The bounding step represents the process of traversing the tree to find the optimal subset [22]. The GA is an adaptive algorithm that finds the global optimum solution for an optimization problem based on the mechanics of natural genetics and biological evolution. GAs operate on a population of individuals to produce better approximations. In attribute selection, each individual in the population represents a predictive model with genes that correspond to the total number of attributes in the dataset. Genes are encoded as binary values that show if the attribute is included or not in the subset [21].

c) *Clustering methods* divide the components of the original set into different groups, and from each cluster, a representative element is selected to compose the optimal subset [24]. The approaches within this category can be further divided into three subcategories: k-means-based [24], [25], affinity propagation-based [26], and graph-based [27].

Among the various subcategories of clustering methods, graph-based clustering methods play a key role. The graph-clustering-based approaches find the relevant attributes by partitioning the graph into subgraphs (clusters) and selecting the representative attribute from each of them [27]. In this representation, the nodes would correspond to data points, while the edge between two nodes is weighted by their similarity. It is important to note that data representation through graphs has attractive characteristics since it enables grasping the local and global properties efficiently. This effect is obtained by the intrinsic ability of graph representation to naturally address local neighborhoods, paths, and global connectivity in its definition [28]–[30]. In this sense, a graph can enhance the characterization of complex manifolds, giving graph-based methods a key role in investigating realistic datasets. Moreover, it can help in reducing the computational complexity of data investigation [27], [30], [31].

When performing dimensionality reduction on graph structures, two main approaches can be addressed. On one side, graph-based clustering algorithms might work on similarities among the nodes according to specific criteria and metrics derived on the attributes associated with each vertex in the graph [27]. Methods belonging to this category (i.e., methods addressing *vertex similarity*) attempt to capture the global geometry of the overall dataset by constructing graphs based on measures of global connectivity of the ensuing graph. The intuition behind this algorithm is that random perturbations of the points in a high-dimensional space will induce changes in a nonhomogeneous fashion in different parts of the graph inducing the given dataset to show minimal global distance. Thus, depending on how globally important certain edges of the graph are, the algorithms working on vertex similarity will aim to capture the globally important edges in the perturbed ensemble [29], [31], [32].

A popular way to ensure such global connectivity addressing vertex similarity is through the minimum spanning tree (MST) approach [27], [33], [34]. The main step of the MST approach consists of determining the MST of the graph, which connects its vertices without cycles and with the minimum total edge weight. MST identifies the graph's cluster by removing the inconsistent edges according to a certain criteria [34]. MST-based approaches can, thus, capture the geometry of nonhomogeneously sampled data points in a high-dimensional space since the MST contains not only local but also global features of the dataset [32], [33], [35].

Another way of performing dimensionality reduction on graphs relies on identifying clusters to fulfill a specific target condition, i.e., a *fitness criterion* [27]. Several forms of fitness criteria have been proposed in technical literature, typically as a function of the density of the clusters to be detected and/or the amount of edges in the graph necessary to reach the maximum value for cliques in the induced subgraphs [27], [36], [37]. In this respect, community detection (CD) algorithms and methods based on dominant set (DS) search play a key role.

Let us consider CD schemes [27], [36], [38], [39]. In general, CD algorithms depend on the definition of the resolution parameter that leads to multiscale CD. Specifically, for small values of this resolution parameter, the number of detected communities is large, and the communities capture the graph's local information. As the resolution parameter becomes larger,

there are fewer communities, and the communities are able to capture the global features of the graph [38], [39]. For instance, Markov stability is a quality measure for CD, which adopts a dynamical perspective to unfold relevant structures in the graph at all scales as revealed by a diffusion process [29], [38], [39].

On the other hand, within fitness-criteria-based graph clustering approaches, DS clustering generalizes the problem of finding a maximal clique to edge-weighted graphs [27], [28], [37]. At each iteration, a DS is extracted, and its subsets of nodes are removed from the graph (this is called the peeling-off strategy). The process iterates on the remaining nodes until all are assigned to a cluster. Hence, the DS approach determines the clusters sequentially using a relative measure that quantifies the clusters' homogeneity [9], [37], [40], [41].

Unlike the ranking and searching algorithms, clustering methods guarantee the nonredundancy of the selected attributes. In this way, the subset of selected attributes is more representative of the original set. Hence, the performance of the remote sensing analysis will be enhanced. Thus, clustering methods, as well as searching methods, are quite accurate. However, graph-based clustering approaches, in particular, are more advantageous than searching methods for their pertinence in dealing with nonconvex datasets. Computational complexity can vary depending on the clustering algorithm that is used and the size of the dataset.

It is worth noting that methods based on deep learning, such as autoencoders, can be used for attribute extraction [42]. By using a training set of data, autoencoders learn a mapping that preserves the structure, from the original data space to a lower dimensional space. Many variants of autoencoders have been proposed for attribute selection as well as to tackle the issue of interpretability loss. Xu *et al.* [43] select the subset of attributes that contributed the most to the output, while Tomar *et al.* [44] backpropagate the network through more probable links, to name a few. The main drawback of approaches based on deep learning is their heavy dependence on the density of the training set. The training dataset should be rich in quality and size to reflect on the structure of the underlying manifold, especially if it has a complex structure. However, due to the difficulty of procuring such dense training sets, such methods can be hardly employed to obtain accurate and reliable results. Moreover, the aforesaid frameworks are not flexible in dealing with heterogeneous datasets. All this adds up to the complexity of implementation. We would like to emphasize that, in this study, we are comparing the proposed method only with unsupervised dimensionality reduction approaches, while neural-network-based approaches are either supervised or semi-supervised; therefore, we are not using any of these approaches since they require a training set [42]–[44].

The methods described above can be classified as supervised if they require labeled data, or unsupervised, otherwise. However, unsupervised methods are more convenient since acquiring labeled data, which in most cases involves the implication of an expert, is costly and time consuming. Indeed, in contrast to other research fields, providing very accurate labels is challenging in the case of remote sensing, for instance, when dealing with complex scenes or when considering modalities that are difficult to interpret, such as SAR images of sea ice in polar areas.

B. Related Work

In this study, we propose an information selection method based on the graph Laplacian. Since this approach has been widely employed for multimodal analysis in remote sensing, it is worth to mention several works based on the graph Laplacian and generally on segmentation of multimodal datasets. The graph Laplacian is a matrix representation of the graph that reflects its properties [9], [12], [45]. In particular, the eigenvectors of the Laplacian constitute a low-dimensional embedding of the nodes (that represent attributes), which increases their separability by revealing their hidden pattern. As opposed to attribute extraction approaches, this embedding can be mapped back, preserving the attributes' physical interpretability. As such, it combines the advantages of attribute selection and attribute extraction methods [17].

The graph Laplacian has been widely applied for multimodal analysis in remote sensing. For instance, we might cite manifold alignment applications that aim to determine a common latent space where multimodal datasets have a unified representation and become comparable [46]. In [47], Tuia *et al.* propose a semi-supervised framework for a manifold alignment that avoids geometric comparisons between modalities since it only compares their labels while preserving each domain's geometry via domain-specific graph Laplacians. A successful outcome of this approach relies on the quality of labels that should be similar among the datasets and representative of their connections. Hong *et al.* [48] propose to consider unlabeled information additionally to labeled samples. In particular, their approach exploits labeled samples from the overlapped area of hyperspectral and multispectral modalities and pseudo labels given only by the multispectral modality. The pseudo labels are updated using a data-driven Laplacian matrix learned on the latent subspaces of both modalities. As opposed to [47], the approach in [48] requires the datasets to be coregistered and overlapped. Furthermore, some deep learning framework attempts to increase the capability of information blending between multimodalities using different strategies, such as multiscale fusion, bidirectional symmetrical mechanism, and highly dense connectivity, have been proposed [49]. Moreover, while, generally, the joint modality representation is used, some methods are building the disjunct subnetworks in order to learn the discriminative features independently for each modality and integrate them with various structured constraints, which can be measured by similarity, correlation, or sequentiality, onto the resulting encoder layers [50]. In addition, some GAN-induced models have also been investigated [51], [52]. Among these methods, it is worth to mention the strategy proposed in [51], where the robustness of the features is increased by eliminating the effects of the adversarial noises. Moreover, the algorithm in [52] models the adversarial perturbation into end-to-end multimodal networks to obtain large-scale semantic segmentation.

Another application is multimodal segmentation, specifically by combining LiDAR and hyperspectral datasets. In [53], Iyer *et al.* proposed an approach based on SC for multimodal segmentation. To combine information from multimodal datasets, the similarity between the pixels is given by the minimum of all

similarities considering different modalities. As such, two classes are similar if and only if they are similar in all modalities. The eigenvectors of the fused graph are then used for segmentation in a semi-supervised manner using the MBO algorithm [53]. Xia *et al.* [54] also propose to combine hyperspectral and LiDAR features in a semi-supervised manner. Their approach exploits both labeled and unlabeled samples to optimally fuse both modalities' spectral, elevation, and spatial features. Hong *et al.* [55] as well as the aforementioned authors further extended their model to a semi-supervised version by learning a graph structure for the alignment of labeled and unlabeled samples. In the case of multimodal datasets that involve the use of information-rich data, which accompanied by high storage and computational costs, it seems relevant to train the model employing only a limited part of the multimodal dataset. Thus, training in more compact and varied cross-modal representations facilitates predicting larger scale semantic segmentation results [55].

For graph building, the GK, also called the heat kernel function or radial basis function (RBF), is typically used to assess the graph's nodes' similarity. In the case of heterogeneous datasets, GK might be a valid choice. However, GK will not be able to reveal the structure of data from different domains [48]. This limitation can be circumvented by comparing the heterogeneous datasets' labels as in [47], assuming that they include similar classes, or by learning the graph from the dataset as in [48]. Both approaches heavily rely on the quality and density of the labels.

C. Contributions

With this in mind, we developed a method for flexible attribute selection based on graph Laplacian representation induced by metrics computed at global and local scales across the given multimodal datasets. When analyzing multimodal data, classical spectral methods are struggling to perform on such highly heterogeneous datasets [10]–[12]. Therefore, in this work, we are suggesting adding another criterion to weight a graph edges in order to solve the limitations of the classic SC approaches. Unlike the commonly used functions to weigh the edges of a graph, such as GK, MI can assess nodes' similarity from different domains since it only compares their probability density functions (PDFs). MI measures the statistical dependence between two random variables. It is defined as the Kullback–Leibler divergence of their joint PDF and the product of marginals. Instead of only exploiting the MI to assess the similarities of multimodal attributes, we propose to combine it with the GK. The GK will compensate for the incapacity of MI to capture the local structure of the attributes. Several works employ two similarity measures for information selection [13], [56]–[58]. In contrast to those methods, we exploit both measures simultaneously and not sequentially. In this way, the results will not be biased by the order in which the measures were applied, i.e., both criteria are equally important, and hence, the selection will be more precise.

Accordingly, while the existing works using graph Laplacians only rely on the attributes' structural similarity using kernels, we also consider the attribute's information content. We jointly employ GK and MI to identify the most relevant attributes within the records. Bearing in mind the variability of the Earth's surface properties, the attributes' relevance will vary among the different classes within the remotely sensed images.

Correspondingly, the second major contribution is that the two similarity measures are applied at different detail levels to preserve more information about original data. The MI is applied globally, i.e., image-wise, so to provide a better estimation of the attributes' shared information. On the other hand, the GK is performed locally, i.e., patch-wise, in order to preserve the structure and particularity of homogeneous areas within the images. This allows us to increase the flexibility and accuracy of information selection since different relevant attributes are selected for various homogeneous areas.

Thereby, the proposed approach guarantees high accuracy of the analysis and reduces the computational complexity so that the potential of multimodal remote sensing data analysis can be exploited in multiple applications. The different experimental tests conducted on several multimodal datasets illustrate the ability of such an approach in revealing the complex pattern of the heterogeneous attributes that ensures a more precise selection than the existing works.

It is worth noting that, as opposed to [53] and [54], our approach employs the Graph Laplacian for attribute selection and not to extract new attributes. As such, we preserve the physical interpretability of the attributes that might be exploited, for instance, in understanding the contribution of each modality in the underlying analysis. Moreover, given the difficulty in acquiring dense and rich labels in remote sensing, and to avoid the imprecision of the selection in case of uncertain labels, our approach is applied in an unsupervised manner.

In order to sum up everything mentioned above, in this article, we introduce an unsupervised, flexible, interpretable, and accurate method for information selection that is applied for multimodal datasets. Among all the mentioned advantages, we would like to stress the main contributions and novelties of this work and proposed approach in particular.

1) *Two Similarities*: It simultaneously employs two similarity measures that preserve global and local particularities of the original dataset, which subsequently allows selecting the most relevant attributes.
2) *Flexible Selection*: The method is performed patch-wise; therefore, it selects the most relevant attributes for the considered classes across the different areas of the ROI.

Additionally, here are some minor advantages, which are less significant, and have been employed in existing works, nevertheless still worth mentioning.

1) *Multimodal*: It is flexible; therefore, it can be applied to various data combinations with different characteristics.
2) *Unsupervised*: The method is completely application independent; thus, it does not require any prior knowledge regarding the datasets or class labels in particular.
3) *Interpretable*: The method keeps the crucial advantages of both dimensionality reduction strategies, namely, attribute extraction and selection, such as preserving the physical meaning of the original data, while increasing its separability.

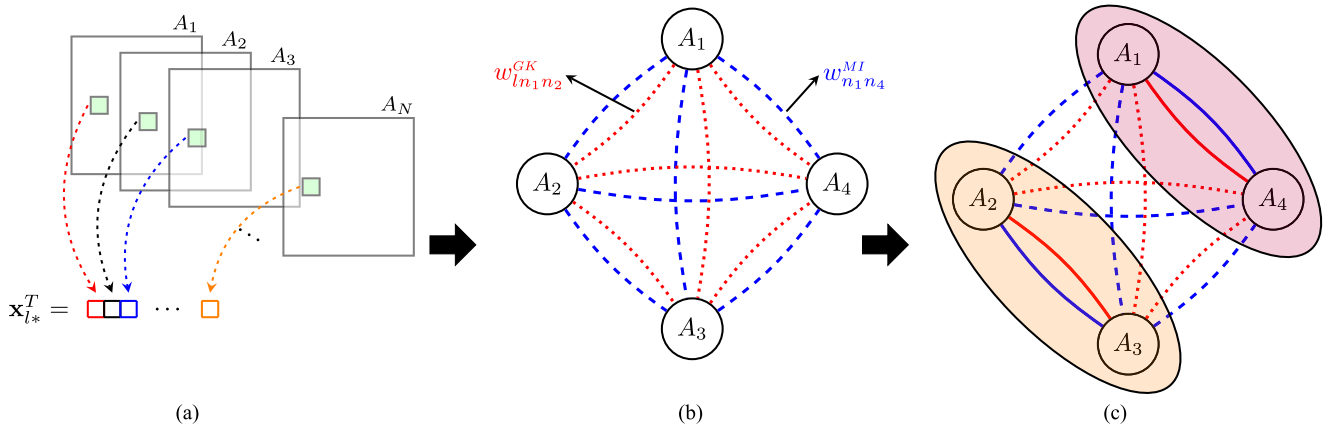


Fig. 1. Flowchart of the multimodal information selection approach proposed in this work. (a) Attributes of the l-th pixel are stacked in one vector $\mathbf{x}_{l*}$. (b) Graph of four attributes with two similarity functions at the l-th pixel. (c) Graph of four attributes with two similarity functions at the l-th pixel. Similar attributes are grouped together.

III. Methods

This section reports the detailed description of the main steps of the proposed information selection method (see Fig. 1).

A. Attribute Generation

The very first step is attribute generation. We assume that the images are spatially aligned such that their attributes overlap. Let M be the number of available images, including bands and polarizations, and L be the number of pixels in each image. We assume that N attributes (images, textural features, etc.) could be associated with each pixel across the whole dataset, and we stack them all in $\mathbf{X} = (x_{ln}) \in \mathbb{R}^{L\times N}$ [see Fig. 1(a)]. We denote the n-th column of $\mathbf{X}$, which corresponds to the n-th attribute by $\mathbf{x}_{*n}$, so it is possible to write $\mathbf{X} = [\mathbf{x}_{*1}, \ldots, \mathbf{x}_{*N}]$. Analogously, we denote the l-th row of $\mathbf{X}$, which details the values of attributes at the l-th pixel, by $\mathbf{x}_{l*}$; hence, $\mathbf{X} = [\mathbf{x}_{1*}^T, \ldots, \mathbf{x}_{L*}^T]^T$.

B. Graph Building

For the sake of clarity, we present our approach first at a pixel level. The adaptation to the superpixel/patch level will be detailed in Section III-D. We aim to find, for a given pixel l, the smallest subset of attributes, $\{x_{l1}, \ldots, x_{lK}\}$, that preserves the structure and information content of the original set. To perform such selection, we apply the graph theory [9] since graphs are a natural way to represent various types of data.

In the proposed method, the set of N attributes will constitute the vertices of an undirected fully connected graph $\mathcal{G}_l(\mathbb{V}_l, \mathbb{E}_l^{\text{GK}}, \mathbb{E}^{\text{MI}})$, where $\mathbb{V}_l$ denote the set of attributes ($\mathbb{V}_l = \{A_1, \ldots, A_N\}$, A_n refers to the n-th attribute the values of which are given by $\mathbf{x}_{*n}$), $\mathbb{E}_l^{\text{GK}}$ and $\mathbb{E}^{\text{MI}}$ are two set of edges that connect the nodes ($\mathbb{E}_l^{\text{GK}}, \mathbb{E}^{\text{MI}} = \{(A_i, A_j), A_i, A_j \subset \mathbb{V}\}$). The weights of the edges are defined by two similarities, GK and MI, to increase the accuracy of analysis (see Fig. 1(b) for an example of four attributes at the l-th pixel).

It is worth noting that two vertices are connected by two edges. The weight of the first edge, between attributes x_{ln_1} and x_{ln_2}, is determined using the GK

$$w_{ln_1n_2}^{\text{GK}} = \exp\left(-\frac{(x_{ln_1} - x_{ln_2})^2}{2\sigma}\right), \qquad 1 \le n_1, n_2 \le N \tag{1}$$

where σ controls the width of the neighborhood in the graph. The width of the neighborhood, i.e., the number of connected vertices, increases with σ. In this work, we set σ to 1 by default, since it produces a more accurate result; however, this parameter does not affect the performance significantly.

A large value of $w_{ln_1n_2}^{\text{GK}}$ implies that the attributes x_{ln_1} and x_{ln_2} are very similar, and hence, it will be sufficient to only consider one of them to obtain accurate characterization of the dataset. Conversely, small values of $w_{ln_1n_2}^{\text{GK}}$ mean that the attributes are different and, therefore, likely to carry different information, so that they must be both considered for the analysis.

The weight of the second edge, between attributes $\mathbf{x}_{*n_1}$ and $\mathbf{x}_{*n_2}$, is defined using MI, as follows:

$$\begin{aligned} w_{n_1,n_2}^{\text{MI}} &= I(\mathbf{x}_{*n_1}, \mathbf{x}_{*n_2}), \qquad 1 \le n_1, n_2 \le N \\ &= \sum_{i=1}^{L}\sum_{j=1}^{L} P(x_{in_1}, x_{jn_2}) \log\left(\frac{P(x_{in_1}, x_{jn_2})}{P(x_{in_1})P(x_{jn_2})}\right) \end{aligned} \tag{2}$$

where $\mathbf{x}_{*n}$ is a vector of measures corresponding to the n-th column of matrix $\mathbf{X}$, i.e., n-th attribute. $P(\mathbf{x}_i, \mathbf{x}_j)$ is the joint density function of $\mathbf{x}_i$ and $\mathbf{x}_j$, and $P(\mathbf{x}_i)$ and $P(\mathbf{x}_j)$ are the marginals. MI quantifies the shared information between two random variables [14]. Accordingly, large values of $I(\mathbf{x}_{*n_1}, \mathbf{x}_{*n_2})$ imply redundancy in information. Conversely, low values of $I(\mathbf{x}_{*n_1}, \mathbf{x}_{*n_2})$ imply synergy (novelty).

The similarity measure based on the GK represents the structure of the attribute set. In our method, it is applied at local level (i.e., on pixels or segments) in order to preserve the local particularities of the original data. On the other hand, MI reports the information content of the attribute set by discarding redundant ones. The selection via information is performed image-wise to capture the global information of the observed region. Thus, we

extract both global and local information about our data in order to enhance the performance of the proposed method.

C. Graph Clustering

Once the graph is defined according to the operations that have been previously introduced, we perform the partition of the graph using a procedure inspired by the SC approach [9] so as to identify and select the most relevant attributes in the dataset. In order to understand the main steps of this strategy, let us suppose that we only use the GK as a similarity measure as is the case in classic SC. The partition is performed by grouping the vertices of the graph into subgraphs so that two vertices of the same subgraph have strong connections (weights), while two vertices from different subgraphs have weak connections. Such a problem can be formalized using the normalized cut criterion [9], which can be defined as follows:

$$\sum_{k=1}^{K} \frac{\sum_{i\in V_{lk}} \sum_{j\in V_l\setminus V_{lk}} w_{lij}^{\mathrm{GK}}}{\sum_{i\in V_{lk}} \sum_{j\in V_{lk}} w_{lij}^{\mathrm{GK}}} \tag{3}$$

where w_{lij}^{GK} is the weight of the edge defined by the GK, and $V_{l1}, \ldots, V_{lK}$ are the K partitions of the graphs, i.e., $\bigcup_k V_{lk} = V_l$. It is also worth recalling that K identifies the number of relevant attributes that are meant to be selected out of the original records. The normalization in (3) ensures that the clusters are large enough to avoid clusters of single vertices. The criterion in (3) is then minimized over the K graph partitions to select the K most relevant attributes in the original dataset.

The aforesaid optimization of the normalized cut criterion is NP-hard and, hence, very cumbersome to efficiently address. To enhance the partition procedure, Shi and Malik proposed to replace the normalized cut minimization with an approximated problem [59]

$$\min_{\mathbf{H}} \mathrm{Tr}\left(\mathbf{H}^T \mathbf{L}_l^{\mathrm{GK}} \mathbf{H}\right) \qquad \text{subject to} \quad \mathbf{H}^T\mathbf{H} = \mathbf{I} \tag{4}$$

where $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K] \in \mathbb{R}^{N\times K}$, and $\mathbf{h}_k$ denotes the indicator vector of the i-th subgraph. $\mathbf{L}_l^{\mathrm{GK}}$ is the so-called symmetric normalized Laplacian matrix based on the GK, and it is defined as follows:

$$\mathbf{L}_l^{\mathrm{GK}} = \mathbf{I} - \mathbf{D}_l^{\mathrm{GK}^{-1/2}} \mathbf{W}_l^{\mathrm{GK}} \mathbf{D}_l^{\mathrm{GK}^{-1/2}} \tag{5}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{W}_l^{\mathrm{GK}} = (w_{lij}^{\mathrm{GK}})$ is the adjacency matrix, and $\mathbf{D}_l^{\mathrm{GK}} = \mathrm{diag}(\sum_{i\neq j} w_{lij}^{\mathrm{GK}})$ is the degree matrix. The n-th element of the graph indicator $\mathbf{h}_k$ can be constrained to assume a nonnull value of $(\sum_{i,j\in V_{lk}} w_{i,j}^{\mathrm{GK}})^{-\frac{1}{2}}$ if and only if the n-th node of the graph belongs to the k-th subgraph.

It is worth noting that such a discrete constraint leads to an NP-hard problem that can be relaxed by allowing the solutions to be in $\mathbb{R}$ [59]. In this case, according to the Rayleigh–Ritz theorem, the solution of (4) is given by the first K eigenvectors of $\mathbf{L}_l^{\mathrm{GK}}$ [9]. In fact, the multiplicity of the null eigenvalue of the Laplacian matrix equals the number of the connected components in the graph, and their corresponding eigenvectors are indicators of different subgraphs [9]. Moreover, since the discrete constraint on the indicators was discarded, a clustering of the rows of $\mathbf{H}$ is required to refine the results [9], [59]. Indeed, the n-th row of $\mathbf{H}$ corresponds to the n-th attribute. As such, the same results of the clustering on the rows of $\mathbf{H}$ apply to the attributes. Moreover, the subset of relevant attributes is constituted by picking, from each cluster, the closest attribute to the centroid. Accordingly, the rows of $\mathbf{H}$ can be considered a revertible low-dimensional embedding of the attributes.

At this point, it is worth recalling that the graph representation of the datasets we aim to analyze is associated with a fully connected graph. In this case, the graph is one connected component. Hence, there will be one null eigenvalue of the Laplacian matrix corresponding to a constant eigenvector [9]. As a consequence, the graph indicators are given by the eigenvectors related to the next lowest eigenvalues. Therefore, graph clustering success relies on the identifiability of these informative eigenvalues related to the graph indicators. As such, they need to be isolated from other eigenvalues [9], [12].

The isolation of the eigenvalues is directly associated with the clusters' separability, which is more plausible in homogeneous graphs, where similar interactions occur among the nodes. On the other hand, the attributes of multimodal datasets are heterogeneous, and they interact differently. In this case, however, it has been shown that the classic graph clustering will fail at separating the clusters [10]–[12]. To tackle this issue, we propose considering the MI in addition to the GK. Incorporating the MI will reflect different relationships between the attributes from the GK. This new variability will help isolate the informative eigenvalues and increases the clusters' separability, which will translate into a precise attributes selection.

Now, if we consider the MI in addition to the GK, we would like to partition the graph such that the vertices of the same subgraph have strong connections via both links, while the vertices from different subgraphs have one or two weak connections, either GK or MI [see Fig. 1(c)]. An approximation of this problem can be written as follows:

$$\begin{cases} \min_{\mathbf{H}} \mathrm{Tr}\left(\mathbf{H}^T \mathbf{L}_l^{\mathrm{GK}} \mathbf{H}\right) \\ \min_{\mathbf{H}} \mathrm{Tr}\left(\mathbf{H}^T \mathbf{L}^{\mathrm{MI}} \mathbf{H}\right) \end{cases} \quad \text{subject to} \quad \mathbf{H}^T\mathbf{H} = \mathbf{I} \tag{6}$$

where $\mathbf{L}^{\mathrm{MI}}$ denotes the Laplacian matrix based on MI

$$\mathbf{L}^{\mathrm{MI}} = \mathbf{I} - \mathbf{D}^{\mathrm{MI}^{-1/2}} \mathbf{W}^{\mathrm{MI}} \mathbf{D}^{\mathrm{MI}^{-1/2}} \tag{7}$$

where the corresponding adjacency matrix and degree matrix are defined as $\mathbf{W}^{\mathrm{MI}} = (w_{ij}^{\mathrm{MI}})$, and $\mathbf{D}^{\mathrm{MI}} = \mathrm{diag}(\sum_{i\neq j} w_{ij}^{\mathrm{MI}})$, respectively. The solution of (6) is given by the common eigenspace of $\mathbf{L}_l^{\mathrm{GK}}$ and $\mathbf{L}^{\mathrm{MI}}$, i.e., their joint eigenvectors. The common eigenspace spanned by both Laplacians enables their interaction, which might unfold complicated structure of the graph. The joint eigenvectors of the graph Laplacians, $\mathbf{L}_l^{\mathrm{GK}}$ and $\mathbf{L}^{\mathrm{MI}}$, are defined so that the following equations hold:

$$\mathbf{L}_l^{\mathrm{GK}} = \mathbf{V}_l \mathbf{\Lambda}_l^{\mathrm{GK}} \mathbf{V}_l^T \tag{8}$$

$$\mathbf{L}^{\mathrm{MI}} = \mathbf{V}_l \mathbf{\Lambda}_l^{\mathrm{MI}} \mathbf{V}_l^T \tag{9}$$

Algorithm 1: SC Algorithm for Local Pixel/Superpixel-wise Selection.

Input:
- Attributes of the l-th pixel—$\{x_{l1}, \ldots, x_{lN}\}$/$l$-th superpixel—$\{\mathbf{x}_{l1}, \ldots, \mathbf{x}_{lN}\}$
- Number of selected attributes—$K < N$

Output: Subset of N Attributes

1) Compute the adjacency matrices $\mathbf{W}^{\text{MI}}$ using (2) and $\mathbf{W}_l^{\text{GK}}$ using (1) for pixel-wise selection and (11) for superpixel-wise selection.
2) Compute the degree matrices $\mathbf{D}_l^{\text{GK}}$ and $\mathbf{D}^{\text{MI}}$
3) Construct the Laplacians $\mathbf{L}_l^{\text{GK}}$ and $\mathbf{L}^{\text{MI}}$ as in (5) and (7), respectively.
4) Compute the first K smallest joint eigenvectors of $\mathbf{L}_l^{\text{GK}}$ and $\mathbf{L}^{\text{MI}}$, $\mathbf{v}_{l1}, \ldots, \mathbf{v}_{lK}$.
5) Form $\mathbf{V}_l = [\mathbf{v}_{l1}, \ldots, \mathbf{v}_{lK}] \in \mathbb{R}^{L \times K}$.
6) Normalize the rows of $\mathbf{V}_l$ to 1.
7) Cluster the rows of $\mathbf{V}_l$ into K clusters using K-means
8) Assign $\mathbf{r}_{li}$ to the same cluster as the i-th row of $\mathbf{V}_l$
9) Return, for each cluster, the closest attributes to the centroid.

where $\mathbf{V}_l = [\mathbf{v}_{l1}, \ldots, \mathbf{v}_{lN}]$ is the matrix of eigenvectors, and $\mathbf{\Lambda}_l^{\text{GK}} = \text{diag}(\lambda_{l1}^{\text{GK}}, \ldots, \lambda_{lN}^{\text{GK}})$ and $\mathbf{\Lambda}_l^{\text{MI}} = \text{diag}(\lambda_{l1}^{\text{MI}}, \ldots, \lambda_{lN}^{\text{MI}})$ are diagonal matrices of the corresponding GK- and MI-based eigenvalues, respectively.

In general, a joint diagonalization (JD) exists if and only if $\mathbf{L}_l^{\text{GK}}$ and $\mathbf{L}_l^{\text{MI}}$ commute in multiplication [60], which is not always valid in practice. Thus, $\mathbf{V}_l$ is determined using approximate JD algorithms [61] instead, which minimize a criterion of diagonality of $\mathbf{V}_l^T \mathbf{L}_l^{\text{GK}} \mathbf{V}_l$ and $\mathbf{V}_l^T \mathbf{L}^{\text{MI}} \mathbf{V}_l$. Different diagonalization constraints and distances can be used leading to a multitude of algorithms. In this work, we perform the JD using the Quasi-Newton algorithm [61], which minimizes the log-likelihood measure introduced by Pham and Cardoso [62], i.e.,

$$\mathcal{L}(\mathbf{V}) = \log \frac{\left|\text{ddiag}\left(\mathbf{V}_l^T \mathbf{L}_l^{\text{GK}} \mathbf{V}_l\right)\right|}{\left|\mathbf{V}_l^T \mathbf{L}_l^{\text{GK}} \mathbf{V}_l\right|} + \log \frac{\left|\text{ddiag}\left(\mathbf{V}_l^T \mathbf{L}^{\text{MI}} \mathbf{V}_l\right)\right|}{\left|\mathbf{V}_l^T \mathbf{L}^{\text{MI}} \mathbf{V}_l\right|}. \tag{10}$$

Once the original set of attributes is embedded into a lower dimensional manifold using the joint null eigenvectors of the Laplacian matrices, a classical clustering method, such as K-means, is then applied to partition the embedding, i.e., to cluster the rows of the matrix $\mathbf{H} = [\mathbf{v}_{l1}, \ldots, \mathbf{v}_{lK}]$ into $K < N$ clusters. This new representation enhances the efficiency of standard clustering methods by increasing the separability of data, mainly if it is nonlinearly separable. Moreover, it eliminates the sensitivity to initialization of such methods. Finally, the centroids of the clusters will form the set of selected attributes. It should be noted that the number of selected attributes K is not determined automatically in this work. Algorithm 1 reports the main steps of the proposed information selection method inspired by Ng *et al.* [63].

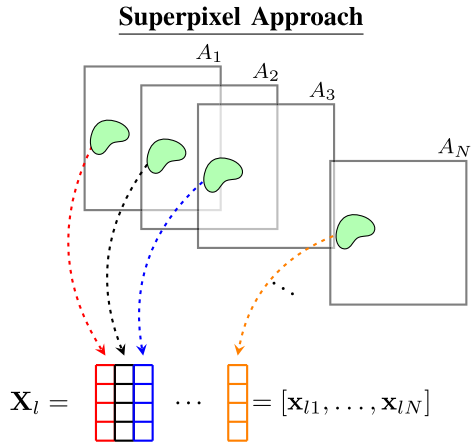


Fig. 2. Attributes of the l-th superpixel are stacked in one matrix $\mathbf{X}_l \in \mathbb{R}^{L_l \times N}$. L_l denotes the number of pixels in the l-th superpixel.

D. Superpixel Approach

Considering the large size of the remote sensing images, performing the selection at each pixel is computationally expensive [64]. To alleviate the computational complexity while preserving each pixel's local particularity, we propose to implement selection on a superpixel-level instead, i.e., patch-wise (see Fig. 2). As opposed to other patch-wise approaches, e.g., windowing, superpixels include pixels that share similar information since they are generated using segmentation (i.e., the grouping of homogeneous pixels) [65]. As such, the selection is more precise since it is particular to the properties of homogeneous pixels.

The first step of a superpixel selection consists of segmenting the image into homogeneous areas. This step can be achieved using segmentation methods such as Watershed [66], [67] or simple linear iterative clustering [68]. In our work, we use Watershed superpixel segmentation.

In the superpixel approach, similar steps as in Algorithm 1 applies except for the calculation of the GK adjacency matrix. In the case of the superpixel-based definition of the graph to be used for attribute selection as previously mentioned in this section, the elements of the adjacency matrix $\mathbf{W}_l^{\text{GK}}$ are calculated using all the pixels within the l-th superpixel, i.e.,

$$w_{ln_1n_2}^{\text{GK}} = \exp\left(-\frac{||\mathbf{x}_{ln_1} - \mathbf{x}_{ln_2}||^2}{2\sigma}\right), \qquad 1 \le n_1, n_2 \le N \tag{11}$$

where $||.||$ denotes the Frobenius norm. The graph is then explored and the eigenanalysis is performed according to the steps detailed in the previous subsection and summarized in Algorithm 1 in order to identify and select the K most relevant attributes in the dataset.

IV. Analysis and Experimental Results

The following section reports the experimental analysis and performance evaluation of the proposed method, as well as comparison results with existing methods using several multimodal

datasets. In the remaining of this section, we refer to our method as GKMI–Gaussian kernel and mutual information.

Attribute selection can be applied as a preprocessing step of several remote sensing applications, e.g., target detection, classification, unmixing, etc. However, for the validation of our method, we only consider the improvement of classification accuracy.

The segmentation step used as part of GKMI might produce superpixels that include a different number of classes, and the classes may differ from one superpixel to another. To tackle the heterogeneity of the superpixels, we classify them separately using parallel classification. Accordingly, we employ L classifiers for the L superpixels that constitute the image $\mathbf{X}$. To train the classifiers, we use the same training set $\mathbb{T} = \{\mathbf{t}_1, \mathbf{t}_2, \dots \mathbf{t}_\theta\}$ that constitutes a certain percentage of the original dataset, where $\mathbf{t}_i \in \mathbb{R}^N$ is the i-th pixel in the training set. However, the attributes of the training set are adapted to each superpixel. As such, for a given superpixel $\mathcal{S}_l$, only a subset of the elements of $\mathbf{t}_i$ is considered. The indices of these elements are the indices of the attributes selected for $\mathcal{S}_l$.

Various classifiers can potentially produce different accuracy results on the same dataset. To validate the performance and show the consistency and robustness of our algorithm, we implement two of the widely applied classifiers in remote sensing: support vector machine (SVM), and random forest (RF) [69], [70].

SVM is a classification method that determines a set of hyperplanes that separate the dataset into different classes [71]. To perform a nonlinear classification, we choose the RBF as a kernel. The optimal parameters c and γ of the RBF kernel are determined by parameter tuning.

RF generates an ensemble of individual decision trees and combines their outputs to get an accurate prediction of the class [72]. In other words, RF is a classifier consisting of a collection of tree-structured classifiers.

Both classifiers are supervised methods that strongly rely on an analyst to define the classes for subsequent classification. To quantitatively estimate the classification result, we use the overall accuracy (OA) index, average accuracy (AA) index, and Cohen's kappa statistic (Kappa). OA shows the percentage of correctly classified samples, AA quantifies the mean of class-specific accuracies for all classes, while Kappa measures the agreement between the classification and the reference data [73].

This section is divided into four subsections, which aim to display the capacity of the proposed method according to the following organization.

1) Section IV-A introduces the datasets that were investigated in this work.
2) Section IV-B investigates the algorithm's sensitivity to the number of selected attributes, the size of superpixels, and the size of the training sample.
3) Section IV-C reports the relevance of using two similarity functions and the pertinence of a superpixel selection versus pixel-wise and image-wise selection.
4) Section IV-D shows the validation of GKMI performance and comparison with different information selection methods on the considered multimodal datasets.

TABLE I
GLCM FEATURES

Features	Definition
Contrast	$\sum_{i,j=0}^{Q-1} g_{i,j}\ (i-j)$
Dissimilarity	$\sum_{i,j=0}^{Q-1} g_{i,j}\ \lvert i-j\rvert$
Homogeneity	$\sum_{i,j=0}^{Q-1} \frac{g_{i,j}}{1+(i-j)^2}$
ASM	$\sum_{i,j=0}^{Q-1} g_{i,j}^2$
Energy	$\sqrt{ASM}$
Correlation	$\sum_{i,j=0}^{Q-1} g_{i,j}\ \left[\frac{(i-\mu_i)(j-\mu_j)}{\sigma_i \sigma_j}\right]$

$g_{i,j}$ denotes the (i,j) element of the GLCM matrix $\mathbf{G}$. Q is the number of gray levels used, and $\mu = \sum_{i=0}^{Q-1}\sum_{j=0}^{Q-1} i\, g_{i,j}$ and $\sigma^2 = \sum_{i=0}^{Q-1}\sum_{j=0}^{Q-1}(i-\mu)\, g_{i,j}^2$ are, respectively, the GLCM mean and variance. ASM refers to the angular second momentum.

TABLE II
TYPES AND NUMBER OF ATTRIBUTES FOR THE CONSIDERED DATASETS

Dataset	Hyperspectral		Optical			LiDAR		N
	Original	GLCM	S2	L8	GLCM	Original	GLCM	
Berlin	–	–	10	18	168	–	–	196
Paris	–	–	10	18	168	–	–	196
Trento	63	–	–		–	2	12	77
Houston	144	–		–	–	1	6	151

S2 and L8 refer to Sentinel-2 and Landsat-8, respectively. N denotes the total number of attributes for each dataset. It should be noted that the GLCM attributes, listed in Table I, are generated for each band.

A. Dataset Description

To evaluate the performance of the proposed GKMI method for attribute selection, we consider different multimodal datasets obtained from various satellite platforms. In this work, we only consider data and feature levels of multimodal data fusion, although the GKMI method can also be applied at the decision level.

To increase the number of attributes and extract some additional information from the original data, along with the bands of optical and LiDAR datasets, we use textural features, while for hyperspectral datasets, we only use existing bands. To extract textural features, we use the gray-level co-occurrence matrix (GLCM) [74]–[76]. Table I illustrates the extracted features as well as their mathematical definitions.

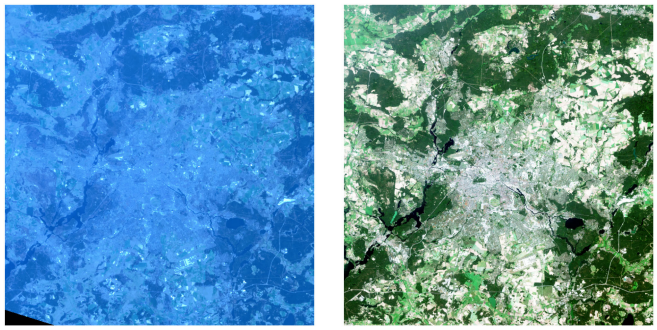
Table II summarizes the number and types of the attributes considered in this article, and Table III reports the list of ground truth labels for each dataset. A detailed description of the datasets is presented as follows.

1) Berlin/Paris: The datasets were acquired over the cities of Berlin and Paris, and both consist of images obtained from two optical sensors: Sentinel-2 and Landsat-8. The datasets were obtained from the 2017 IEEE GRSS Data Fusion Contest [77].

Both datasets (Sentinel-2 and Landsat-8) were resampled at 100-m resolution. Berlin and Paris test sites were prelabeled for the subsequent classification and include 12 ground truth labels corresponding to various built-up (anthropogenic constructions)

TABLE III
GROUND TRUTH LABELS FOR ALL THE DATASETS USED IN THIS ARTICLE

	Berlin	Paris	Trento	Houston
ω_1	Compact midrise	Compact high-rise	Buildings	Grass healthy
ω_2	Open high-rise	Compact midrise	Wood	Grass stressed
ω_3	Open midrise	Open high-rise	Apple trees	Grass synthetic
ω_4	Open low-rise	Open midrise	Roads	Tree
ω_5	Large low-rise	Open low-rise	Vineyard	Soil
ω_6	Sparsely built	Large low-rise	Ground	Water
ω_7	Dense trees	Sparsely built		Residential
ω_8	Scattered trees	Dense trees		Commercial
ω_9	Bush and scrub	Scattered trees		Road
ω_{10}	Low plants	Low plants		Highway
ω_{11}	Bare soil or sand	Bare rock or paved		Railway
ω_{12}	Water	Water		Parking lot 1
ω_{13}				Parking lot 2
ω_{14}				Tennis court
ω_{15}				Running track



(a) (b)

Fig. 3. Overlapping area of the Berlin dataset. (a) Landsat-8 and (b) Sentinel-2 natural color composite images.

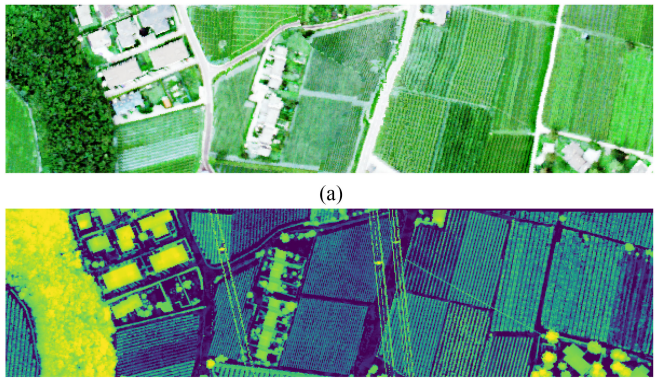
(a) (b)

Fig. 4. Overlapping area of the Paris dataset. (a) Landsat-8 and (b) Sentinel-2 natural color composite images.

and land cover types. Sentinel-2 dataset contains ten bands in the visible, near-infrared, and short-wave infrared part of the spectrum. Landsat-8 contains nine bands in visible, short, and long infrared wavelengths (according to the notation in Section III-A, $M = 28$, 1×Sentinel-2 dataset + 2×Landsat-8 datasets). Moreover, from each band, we extract six textural features (see Table I). Therefore, the final datasets that were used contain $N = 196$ extracted attributes.

Figs. 3 and 4 show the overlapping area of the two datasets. The overlapping test size area for Berlin example is of 666×643 pixels, and for Paris, it is of 988×1160 pixels.

2) Trento: This dataset was acquired on an agricultural area in the south part of the city of Trento, Italy. It consists of LiDAR and hyperspectral data. Hyperspectral data were acquired by the



(a)

(b)

Fig. 5. False-color composite representation of Trento (a) hyperspectral and (b) LiDAR datasets.

AISA Eagle sensor with a 1-m spatial resolution and includes 63 bands ranging from 0.40 to 0.99 μm, where the spectral resolution is 9.2 nm. The LiDAR data were acquired by the Optech ALTM 3100EA sensor. The available ground truth labels consist of six classes.

The Trento dataset contains 63 hyperspectral bands and two LiDAR bands ($M = 65$). Additionally, we extracted six textural features for each of the available LiDAR bands (see Table I). The final dataset that was used contains $N = 77$ attributes with an overlapping test size area of 600×166 pixels. Fig. 5 illustrates the false-color composite representation of the Trento dataset for both sensors.

3) Houston: The last dataset consisted of LiDAR and hyperspectral data acquired over the University of Houston campus and the neighboring urban area and was distributed for the 2013 IEEE GRSS Data Fusion Contest [78]. Hyperspectral data were acquired from the Compact Airborne Spectrographic Imager with a 2.5-m spatial resolution. The hyperspectral dataset includes 144 spectral bands ranging from 0.38 to 1.05 μm. The available ground truth labels consisted of 15 classes.

The Houston dataset contains hyperspectral data (144 bands) and 1×LiDAR data (including one band and six textural features). The final dataset that was used consisted of $N = 151$ attributes with an overlapping test size area of 1905×349 pixels. Fig. 6 demonstrates the Houston test site for both sensors.

B. Parameter Sensitivity Analysis

Several parameters may affect the performance of GKMI, mainly the number of selected attributes, the size of superpixels, and the size of the training set. In the following, we tune one parameter at a time to understand how it influences our proposed approach.

1) Number of Attributes: Fig. 7 illustrates the overall accuracies for the proposed GKMI attribute selection over a different number of selected attributes for all datasets used in this work. The blue curve identifies the OA results obtained on the Berlin dataset, the red line indicates the OA results obtained on the Paris dataset, the green line refers to the Trento dataset, whereas the black line shows the results for the Houston dataset. The stars

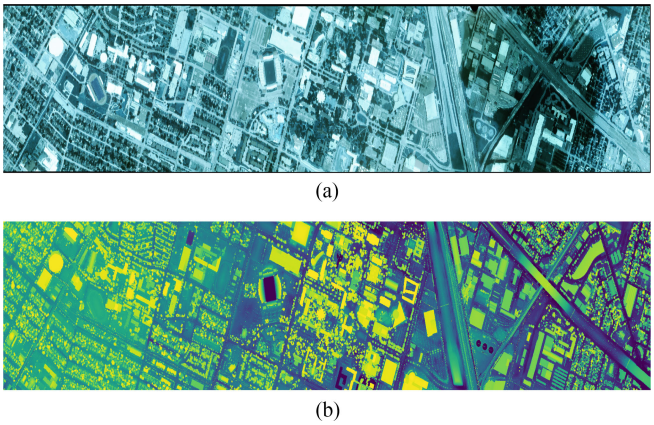


Fig. 6. False-color composite representation of Houston (a) hyperspectral and (b) LiDAR datasets.

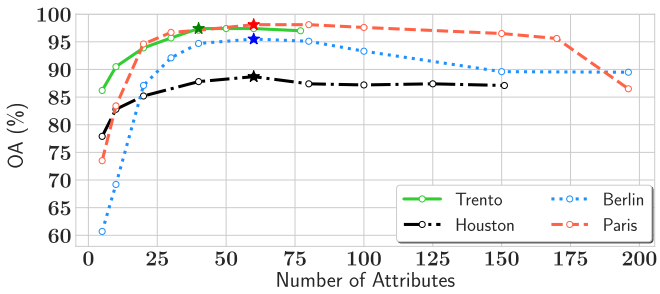


Fig. 7. Overall accuracies of GKMI as a function of different numbers of selected attributes for Berlin (blue dotted line), Paris (red dashed line), Trento (green solid line), and Houston (black dash-dotted line) test sites using the SVM classifier.

indicate the point where the accuracy reaches its maximum. It can be seen from Fig. 7 that the OA curves rise sharply until the number of attributes chosen reaches 40 for the Berlin dataset and 30 for the Paris dataset. After that point, OA curves keep stable high till 75 and start to decrease. The Trento and Houston curves grow abruptly until the number of attributes chosen reaches 20 for both datasets.

All the curves, in general, have a similar pattern that indicates that a large number of selected attributes do not necessarily lead to the best classification result. The number of selected attributes reaches some particular point where additional attributes can hardly provide any extra information for subsequent classification. Depending on the original data, additional attributes may even reduce the accuracy of classification. This result shows the relevance of our method, since using the total number of attributes leads to lower accuracy. Actually, the maximum efficiency is reached using less than half of the attributes.

2) Size of Superpixels: The size of superpixels is another parameter that may impact the performance of our method. Since the same set of attributes is assigned to the pixels of the same superpixel, we expect that too large or too small superpixels may deteriorate the results. Large superpixels may include several homogeneous regions; hence, the selected subset may not be representative of all pixels. On the other hand, small superpixels

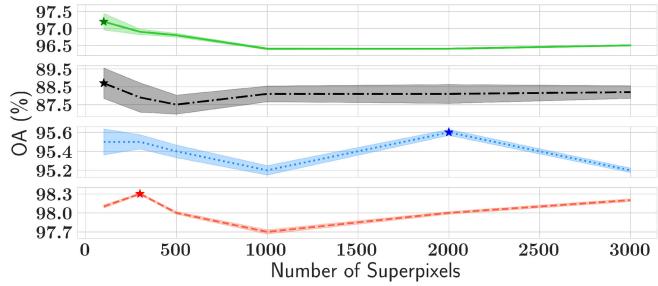


Fig. 8. Mean (curves) and variance (faded surfaces) of overall accuracies of GKMI as a function of different numbers of superpixels for Berlin (blue), Paris (red), Trento (green), and Houston (black) test sites using the SVM classifier. The same color legend as in Fig. 7 applies here. Note that the figure consists of four different subfigures with different scales on the vertical axes.

may not contain the whole homogeneous area, and various attributes can be chosen for the same region.

It is important to recall that the superpixel selection is used only during the attribute selection process, while classification is performed for each pixel separately. In other terms, let us assume that the i-th superpixel $\mathcal{S}_i$ consists of P_i pixels $\{\mathbf{x}_{p*}\}_{p=1,\ldots,P_i}$. The pixels belonging to $\mathcal{S}_i$ cannot be associated with any other superpixel, i.e., $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset \, \forall (i,j) \in \{1,\ldots,L\}^2, i \neq j$, where L is the total number of the considered superpixels in the dataset. Then, the attribute selection procedure in Section III selects for all the pixels in $\mathcal{S}_i$ a subset of K attributes Ω_i. The p-th pixel in $\mathcal{S}_i$ is, hence, classified independently from the others by taking into account only the attributes in Ω_i.

To investigate the impact of this parameter on GKMI, we illustrate in Fig. 8 the OA of the proposed method, over a different number of superpixels, for all datasets. The blue line shows the OA result for Berlin, the red line for Paris, the black line for Houston, and the green line for Trento. The stars show the point with maximum OA. The faded area displays the variance of the overall accuracies for different sizes of superpixels. The number of superpixels is representative of the size of superpixels in the dataset, i.e., the higher the number of superpixels, the smaller the size of the superpixels.

From Fig. 8, we can observe that the curves for each dataset are quite stable, and there are no significant fluctuations, which means that the size of the superpixels has a minor impact on the classification accuracy. Moreover, from the curves, it is possible to appreciate that the variance of the overall accuracies (faded area) is decreasing with the size of the superpixels for the Trento, Houston, and Berlin datasets. This indicates that for these particular examples, increasing the number of superpixels leads to a more stable results, while for the Paris dataset, there are no significant fluctuations in variance throughout the curve.

The OA displayed in Fig. 8 is the result of the attribute selection process as a function of the number of superpixels (L in the previous discussion). Therefore, it is possible to state that Fig. 8 shows how robust the proposed method is with respect to the L parameter. In fact, although the pixels of a given superpixel have the same set of chosen attributes, they might belong to different classes. Thus, the proposed approach is able to combine the benefits provided by the superpixel grouping

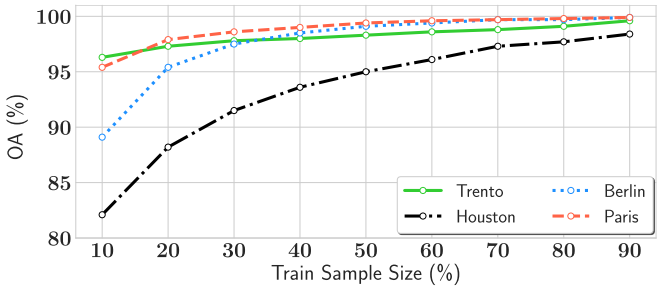


Fig. 9. Overall accuracies of GKMI as a function of different sizes of training samples for Berlin, Paris, Trento, and Houston test sites using the SVM classifier. The same color legend as in Fig. 7 applies here.

and graph clustering while avoiding biasing the results in the classification step.

3) Size of Training Sample: Another parameter that is affecting the performance is the size of the training sample. Fig. 9 demonstrates the OA of the proposed method over a different size of training samples. It is quite evident from Fig. 9 that the increase in accuracy is directly proportional to the increase in the sample size: this behavior is verified for all considered datasets.

In all other experiments of this study, we are using 20% of ground truth labels as a training sample as commonly used in practice.

C. Performance Analysis

Let us now investigate the impact of the chosen similarity measures, their weight, as well as the relevance of the superpixel analysis compared to pixel-wise and image-wise selection.

1) Kernel Comparison: Let us start by discussing the metric to be used to estimate the similarity representing the structure of the attributes' set in Section III. In this respect, it is worth noting that the choice of the function to model the similarity among attributes is a critical task in remote sensing data analysis [79], [80, ch. 9].

In fact, defining the kernel to be employed to quantify the structure of the data (and thus defining similarity between pairs of samples) is crucial to obtain a reliable understanding of the relevance of the attributes and their actual role in the characterization of the interactions among the records [79]. Furthermore, it is important to recall that a proper choice of the kernel to quantify the similarity among attributes can provide a consistent and well-founded theoretical framework for developing nonlinear techniques. Moreover, kernel functions are used in practice to unfold the complicated structure of a dataset, thus enabling the ability to deal with a low number of (potentially high dimensional) training samples, the investigation of heterogeneous records, as well as considering multiple noise sources [79], [81].

On the other hand, it is also true that the definition of the proper kernel for the aforesaid purpose might be particularly cumbersome, especially when the design of architectures for data analysis that is intended to be unsupervised, versatile, and flexible is targeted [79], [81]. Indeed, it is possible to state that the definition of a kernel mapping function that would accurately quantify the similarity among samples represents a bottleneck

TABLE IV
DEFINITIONS OF THE DIFFERENT KERNELS BETWEEN FIRST AND SECOND ATTRIBUTES IN THE l-TH SUPERPIXEL

	Definition	Parameters
Euclidean distance (ED)	$\|\|\mathbf{x}_{ln_1} - \mathbf{x}_{ln_2}\|\|$	–
Linear kernel (LK)	$\mathbf{x}_{ln_1}^T \mathbf{x}_{ln_2} + c$	c: Optional constant
Polynomial kernel (PK)	$\left(s\mathbf{x}_{ln_1}^T \mathbf{x}_{ln_2} + c\right)^d$	c: Optional constant d: Polynomial degree s: Slope
Gaussian kernel (GK)	$\exp\left(-\frac{\|\|\mathbf{x}_{ln_1} - \mathbf{x}_{ln_2}\|\|^2}{2\sigma}\right)$	σ: scale parameter

TABLE V
PERFORMANCE COMPARISON AMONG DIFFERENT SIMILARITY MEASURES, DEFINED IN TABLE IV, USING THE SVM CLASSIFIER

	Berlin		Paris		Trento		Houston	
Method	K	OA (%)	K	OA (%)	K	OA (%)	K	OA (%)
ED	100	93.8	80	97.7	60	97.2	100	87.3
LK	100	88.0	100	96.3	60	97.1	100	87.6
PK	100	90.2	100	96.2	60	97.2	125	86.8
GK	80	**94.9**	80	**97.9**	60	**97.6**	80	**87.9**

K refers to the optimal number of selected attributes for which the OA is obtained. In this experiment, according to the notation in Table IV, $c = 1$, $d = 3$, and $s = 1/N$.

for any kernel-based analysis approach. At the same time, it is worth remembering that not all kernel similarity functions are allowed. Specifically, valid kernels must fulfill Mercer's theorem, i.e., being positive-definite similarity matrices. This property is fundamental when no *a priori* knowledge on the interclass and intraclass statistical distributions is available [79], [81]. As a result, the kernel functions that are most commonly employed in this context are using Euclidean distance (ED) and linear, polynomial, and Gaussian functions (i.e., Linear kernel (LK), Polynomial kernel (PK), and GK in Table IV) as similarity measures.

Thus, to assess the relevance of the choice we proposed in Section III, we compare the performance of the SC in (4) when using ED, LK, PK, and GK to define the weights of the graph structure. In Table V, we represent the maximum OA achieved by the attributes selected using the different kernels. It can be seen that ED, PK, and LK show a slightly lower accuracy than GK. Moreover, the GK always achieves the highest OA with a fewer number of attributes. Compared to ED, LK, and PK, the GK is able to unfold the finer structure of the attributes since it is highly nonlinear. This strengthens the assumptions we have used in designing the data analysis steps in the proposed multimodal feature selection method. This result is consistent with the proven ability of GK to be more flexible in characterizing the data structure in complex systems, especially when an investigation of large-scale and heterogeneous datasets is conducted [45].

2) Significance of Similarity Functions: It is also worth to investigate the relevance of using MI together with the GK similarity to build the graph representing the structure of the dataset. According to the assumptions we have detailed in Section III, the proposed method assumes that both the GK and the MI are

necessary to obtain a solid characterization of the data structure to be analyzed. Indeed, the graph representation plays a key role in describing the interactions among attributes [9], [30]. Therefore, exploring the impact of the chosen similarity metrics in the definition of the graph induced by the considered dataset is crucial to understand what role the quantities used to describe the attributes' relevance can play in different applications, as well as to estimate the reliability of the proposed approach in operational use.

Let us then investigate the impact of GK and MI in the selection process outlined in Section III. In particular, to this aim, the weight of MI and GK metrics could be, in principle, unevenly distributed. Specifically, we can rewrite the function in (10) as follows:

$$\mathcal{L}(\mathbf{V}) = \left(\alpha \log \frac{\left|\text{ddiag}\left(\mathbf{V}_l^T \mathbf{L}_l^{\text{GK}} \mathbf{V}_l\right)\right|}{\left|\mathbf{V}_l^T \mathbf{L}_l^{\text{GK}} \mathbf{V}_l\right|} + (1-\alpha) \log \frac{\left|\text{ddiag}\left(\mathbf{V}_l^T \mathbf{L}^{\text{MI}} \mathbf{V}_l\right)\right|}{\left|\mathbf{V}_l^T \mathbf{L}^{\text{MI}} \mathbf{V}_l\right|} \right). \quad (12)$$

In other terms, the parameter α is used to change the weight (i.e., importance) of the similarity metrics employed in the selection process, i.e., high values of α give more weight to the GK, while low values of α give more weight to the MI. Particularly, only MI is considered when $\alpha = 0$, and only the GK is utilized when $\alpha = 1$.

2) Eigenvalue analysis: At this point, we study the impact of the two similarity metrics on classification performance by investigating the spectrum of the eigenvalues for different values of α. In fact, as previously mentioned in Section III-C, the eigenvectors of the Laplacian matrices used to describe the graph connectivity induced by the given dataset are directly linked to the solution of the feature selection process itself. Indeed, it is worth recalling that the key idea of graph clustering based on Laplacian matrices is that the indicators of data (attributes) classes are given by the eigenvectors of the Laplacian corresponding to the lowest eigenvalues [63]. Furthermore, heterogeneity in the graph node degrees would translate in spreading the eigenvalues of the Laplacian matrix across the spectrum [12]. This means that in the case of complex datasets (i.e., datasets where it is not possible to draw linear hyperplanes in the attribute space to perform graph clustering and therefore dimensionality reduction), it is not possible to associate the informative eigenvectors with the smallest eigenvalues anymore. Actually, by losing the isolation of informative eigenvalues, the associated eigenvectors tend to merge with the eigenvectors associated with close-by (noninformative) eigenvalues [12]. Hence, an effective dimensionality reduction can be performed only when it is possible to identify the smallest eigenvalues and clearly separate them by the eigenvalues with higher amplitude. On the contrary, when the spectrum of the eigenvalues is generally flat, then it is possible to expect that the dimensionality reduction process would not be able to achieve reliable and robust results in terms of informativity maximization [12], [63].

With this in mind, we computed the eigenvalues of the Laplacian matrix resulting from setting α to several values in [0,1]. We then considered their spectra to understand how easy it would be to identify and discriminate the smallest eigenvalues from their total set. In this respect, Fig. 10 shows the eigenvalues of the Laplacian matrices L^{GK} (case $\alpha = 1$ in (12)—yellow solid line) and L^{MI} (case $\alpha = 0$ in (12)—red dotted line) defined in (5) and (7), respectively, as well as their common eigenvalues used by GKMI (case $\alpha = 0.5$ in (12)—blue dashed curve) obtained on Paris and Trento datasets.

We notice that the curves corresponding to the eigenvalues of $\mathbf{L}^{\text{GK}}$ and $\mathbf{L}^{\text{MI}}$ are essentially flat [see, for instance, the enlarged section of the graph on the GK eigenvalues' trend in Fig. 10(a)]. The amplitude of the eigenvalues varies in both cases in the order of 10^{-12}), showing a low separability of the data since it is hard to isolate the eigenvalues related to class indicators [12]. On the other hand, when using both similarities according to the proposed method in Section III, it is possible to appreciate that the variability of the eigenvalues' amplitude is more pronounced in terms of several orders of magnitude. Therefore, it is a lot easier to identify the smallest eigenvalues and separate them from the total set of eigenvalues, leading to a more accurate identification of the relevant attributes in the dataset.

Fig. 10, thus, demonstrates how the heterogeneity of multimodal attributes makes their structure so complex in the attribute space such that the classic SC fails to reveal it. It is indeed worth noting that this result is compliant with recent findings in technical literature, where it has been shown that SC fails at detecting the classes of a graph with heterogeneous degrees [10]–[12], as it is the case in this work.

2) Impact of α on OA: The aforesaid results are confirmed when exploring the classification accuracy obtained when dimensionality reduction is performed for different values of α. Specifically, Fig. 11 shows the gain in the OA of GKMI compared to SC [i.e., $\alpha = 1$ in (12)] as a function of the parameter α. We notice that a negative gain (loss) is only achieved for $\alpha = 0$, implying that MI shows lower performance than the GK and that the exploitation of both measures always improves the OA. The maximum accuracy is achieved when both GK and MI are employed to define the graph [i.e., $\alpha \neq \{0, 1\}$ in (12)], specifically for $\alpha = 0.7$ for the Berlin dataset and $\alpha = 0.5$ for other datasets. According to the trends in Fig. 11, $\alpha = 0.5$ seems the best design choice to achieve high accuracy performance while guaranteeing wide applicability of the system to data with different properties.

2) Impact of α on K: To further demonstrate the pertinence of using two similarities, we compare the results obtained when using one similarity at a time and when used together. Table VI demonstrates the maximum OA of SVM classification for Berlin, Paris, Trento, and Houston datasets, when the selection is performed using only GK, using only MI, and with both similarities (GKMI). In contrast to Fig. 11, where a fixed number of chosen attributes are used, Table VI shows the optimal number for which the maximum OA is achieved. The different approaches show almost similar performance. However, GKMI reaches the maximum accuracy with less number of attributes for each dataset. It achieves an OA of 95.5% for Berlin and 98.1% for Paris, for less than a third of the original attribute set. Moreover, it achieves an OA of 88.7% for Houston, with less

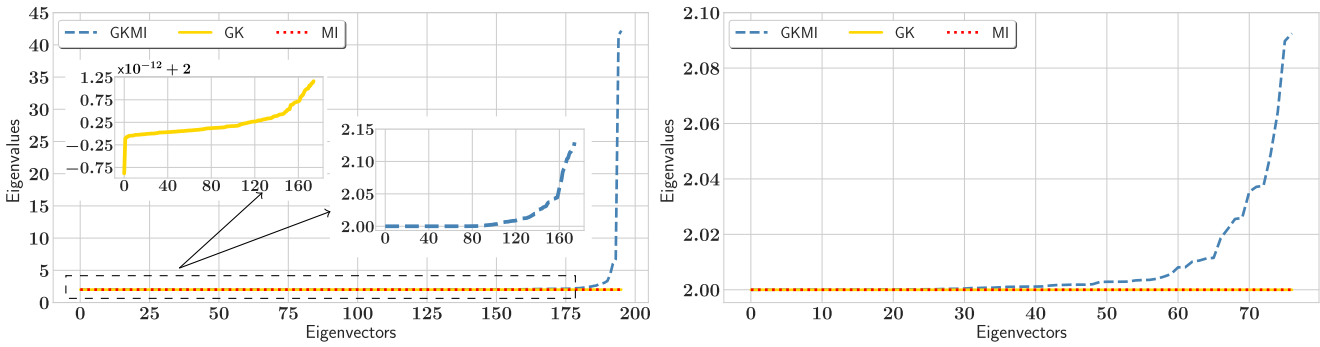


Fig. 10. Amplitude of the eigenvalues of different Laplacian matrices for (a) Paris and (b) Trento datasets as a function of the number of eigenvectors (directly linked to the number of clusters that can be drawn in graph partitioning [9], [12]). The amplitude curves associated with eigenvalues of $\mathbf{L}^{\text{GK}}$ [as for (5)—case $\alpha = 1$ in (12)] and the eigenvalues of $\mathbf{L}^{\text{MI}}$ [as for (7)—case $\alpha = 0$ in (12)] are plotted in yellow solid line and red dotted line, respectively. The amplitude of the eigenvalues obtained via JD [proposed approach in Section III—case $\alpha = 0.5$ in (12)] is displayed in dashed blue line.

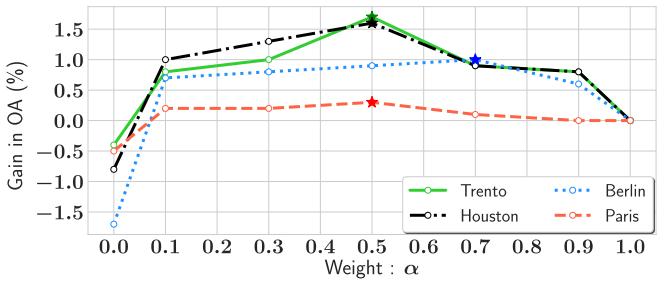


Fig. 11. Gain in the overall accuracies of GKMI compared to SC over a different values of α for Berlin (blue), Paris (red), Trento (green), and Houston (black) test sites using the SVM classifier. The same color legend as in Fig. 7 applies here.

TABLE VI
PERFORMANCE COMPARISON OF SINGLE AND JOINT SIMILARITY MEASURES, USING THE SVM CLASSIFIER

	Berlin		Paris		Trento		Houston	
Method	K	OA (%)	K	OA (%)	K	OA (%)	K	OA (%)
MI	80	93.7	80	97.5	60	96.8	80	86.6
GK/Spectral clustering	80	94.9	80	97.9	60	**97.6**	80	87.9
GKMI	**60**	**95.5**	60	**98.1**	**40**	97.4	**60**	**88.7**

K refers to the optimal number of selected attributes for which the maximum OA is obtained.

than half of the dataset. For the Trento dataset, the OA reaches the highest number 97.4% with 40 attributes, almost half of the dataset. Accordingly, we can conclude that GKMI ensures a more precise selection.

3) Fusion of Similarity Functions: For a given superpixel l, GKMI generates a graph with two edges that summarize the attributes' similarities via the GK and via MI. The two edges are then combined in a nonlinear manner by extracting the joint eigenspace of their corresponding Laplacians, $\mathbf{L}_l^{\text{GK}}$ and $\mathbf{L}^{\text{MI}}$. The fusion of the edges can be performed differently. One of the easiest approaches is by taking their mean

$$\mathbf{W}_l^{\text{mean}} = \frac{1}{2}\left(\mathbf{W}_l^{\text{GK}} + \mathbf{W}^{\text{MI}}\right). \tag{13}$$

In this case, the indicators of the subgraphs are given by the first K eigenvectors of the Laplacian matrix

$$\mathbf{L}_l^{\text{mean}} = \mathbf{I} - \mathbf{D}_l^{\text{mean}^{-1/2}} \mathbf{W}_l^{\text{mean}} \mathbf{D}_l^{\text{mean}^{-1/2}} \tag{14}$$

where $\mathbf{D}_l^{\text{mean}} = \text{diag}(\sum_{i \neq j} w_{lij}^{\text{mean}})$.

Another approach of graph fusion was proposed by Iyer *et al.* [53]. Their approach assumes that two nodes are similar if and only if they are similar via both similarity functions. As such, they define the weight of the combined edge as the maximum of both edges normalized

$$\mathbf{W}_l^{\max} = \max\left(\frac{\mathbf{W}_l^{\text{GK}}}{\text{std}(\mathbf{W}_l^{\text{GK}})}, \frac{\mathbf{W}^{\text{MI}}}{\text{std}(\mathbf{W}^{\text{MI}})}\right) \tag{15}$$

where $\text{std}(\mathbf{A})$ denotes the standard deviation of the elements of the matrix $\mathbf{A}$. In this case, as for the "mean" approach, the indicators of the subgraphs are given by the first K eigenvectors of the Laplacian matrix

$$\mathbf{L}_l^{\max} = \mathbf{I} - \mathbf{D}_l^{\max^{-1/2}} \mathbf{W}_l^{\max} \mathbf{D}_l^{\max^{-1/2}} \tag{16}$$

where $\mathbf{D}_l^{\max} = \text{diag}(\sum_{i \neq j} w_{lij}^{\max})$.

To evaluate the performance of the different approaches, we compare their spectra to assess their ability in separating the different classes of the heterogeneous attributes.

Fig. 12 shows the eigenvalues of the Laplacian matrices obtained by the *"mean"* approach, $\mathbf{L}_l^{\text{mean}}$, the *"max"* approach, $\mathbf{L}_l^{\max}$, and by the *"joint"* decomposition used by the GKMI approach. The flatness of the curves corresponding to *"mean"* and *"max"* demonstrates the incapacity of these approaches in emphasizing the informative eigenvalues, corresponding to the eigenvectors indicators of the attributes' clusters, since they are inseparable from the total set of eigenvalues. On the other hand, when using GKMI, the informative eigenvalues are well isolated. This outcome demonstrates the effectiveness of our approach.

In fact, the JD of the Laplacian matrices corresponding to the GK and MI enables their interaction, revealing their nonlinear connections and, hence, the hidden structure of the heterogeneous attributes. By connection between the similarity functions, we mean the connection of the graph structures that each

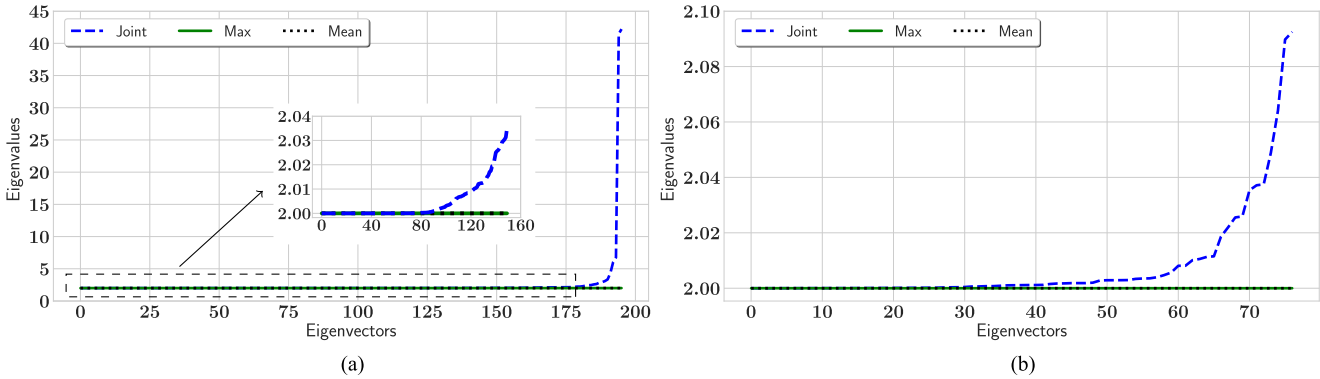


Fig. 12. Amplitude of the eigenvalues of the different approaches to combine the similarity functions for (a) Paris and (b) Trento datasets as a function of the number of eigenvectors (directly linked to the number of clusters that can be drawn in graph partitioning [9], [12]). The amplitude curves associated with eigenvalues of $\mathbf{L}^{\text{mean}}$ (14) and the eigenvalues of $\mathbf{L}^{\max}$ [proposed in [53] and defined in (16)] are plotted in black dotted line and green solid line, respectively. The amplitude of the eigenvalues obtained via JD [proposed approach in Section III in (10)] is displayed in dashed blue line.

TABLE VII
PERFORMANCE COMPARISON OF PIXEL, SUPERPIXEL, AND IMAGE-WISE GKMI APPROACHES FOR EACH DATASET

Dataset	K	Approach	OA (%) SVM	OA (%) RF	ET (sec)
Berlin	40	Pixel-Wise	19.9	24.2	44320
		Image-Wise	94.2	91.6	394
		Superpixel	**94.7**	**93.5**	**246**
	N = 196	No selection	89.5	93.3	465
Paris	40	Pixel-Wise	38.2	51.4	52880
		Image-Wise	96.8	94.6	889
		Superpixel	**97.1**	**95.3**	**267**
	N = 196	No selection	86.5	95.4	758
Trento	40	Pixel-Wise	55.2	66.6	3900
		Image-Wise	96.4	96.4	**23**
		Superpixel	**97.4**	97.3	64
	N = 77	No selection	97.0	**97.7**	26
Houston	40	Pixel-Wise	15.7	32.3	35310
		Image-Wise	84.3	81.0	296
		Superpixel	**87.8**	**85.8**	81
	N = 151	No selection	87.1	85.5	**58**

For the superpixel part of the calculations, 100 superpixels were used.

similarity represents. Conversely, the mean and max approaches assume linear and simple links between the similarity functions, which fail to identify and characterize their nonlinear links, and hence do not exploit their full potential.

4) Levels of Spatial Detail: The proposed GKMI method can be applied at different fusion levels. Nonetheless, its versatility allows us to investigate its application at different spatial detail levels, as GKMI can run at an image, superpixel, and pixel levels. Each of these approaches produces a different result in terms of classification accuracy and time complexity. Therefore, it is interesting to investigate how this design choice might affect the final outcome of the attribute selection procedure.

Table VII shows the OA and execution time (ET) for image, pixel, and superpixel GKMI on all datasets used in this work. It is clear from these results that the superpixel method produces higher accuracy outcomes and outperforms pixel and image-wise approaches for different classifiers, in terms of both accuracy and computational complexity. Furthermore, the ET can be further enhanced by applying parallel computing on the different superpixels.

The superpixel procedure accounts for the particularity of each superpixel, in contrast to the image-wise, and selects the same attributes for homogeneous regions, as opposed to pixel-wise. These two reasons make the superpixel approach more accurate and effective.

Let us now investigate in more detail the GKMI approach performed at the superpixel level. In this case, the adjacency matrix using the GK is measured using all pixels of a given superpixel, as shown in (11). However, in view of the fact that the superpixels in our analysis are formed by grouping homogeneous pixels, we can improve our analysis' time complexity by performing the selection by considering each attribute's mean over all pixels or by picking a representative pixel randomly. As such, for a given superpixels with L pixels, the input of Algorithm 1 for attribute selection is a set of scalars instead of vectors given by the mean of the attributes $\{\frac{1}{L}\sum_l \mathbf{x}_{l1}, \ldots, \frac{1}{L}\sum_l \mathbf{x}_{lN}\}$ or by the attributes of the l-th randomly picked pixel $\{x_{l1}, \ldots, x_{lN}\}$. The algorithm's output, i.e., the subset of relevant attributes, will then be applied to all pixels within the superpixel. Table VIII shows the comparison of these approaches for Berlin, Paris, Trento, and Houston datasets. The results show that by randomly picking a representative pixel, the time complexity reduces without significantly affecting the OA.

In order to strengthen the idea and motivation behind the employment of the information selection on a superpixel level, we additionally analyzed the attributes that were selected by the proposed method for each class of the Trento dataset. As was mentioned earlier, the Trento dataset consists of 77 attributes (63 hyperspectral bands ranging from 402.89 to 989.09 nm, and 14 LiDAR + GLCM textural features) and has six ground truth classes, including Apple trees, Vineyard, Wood, Roads, Ground, and Buildings. Accordingly, Fig. 13 illustrates the chord diagrams that represent selected attributes for five different superpixels that fall into the area of the ground truth labels for each class of the Trento dataset. The vertices show 77 available

TABLE VIII
OA AND ET OBTAINED WITH THREE SUPERPIXEL SELECTION APPROACHES

	Berlin			Paris			Trento			Houston		
	OA			OA			OA			OA		
	μ	σ^2	ET	μ	σ^2	ET	μ	σ^2	ET	μ	σ^2	ET
All pixels	95.5	0.11	359	**98.1**	0.03	336	**97.4**	0.29	**68**	**88.7**	0.71	88
Mean of the pixels	**95.7**	0.06	311	98.0	0.02	279	95.9	0.52	**68**	87.6	0.61	73
One pixel randomly picked	95.5	0.19	**309**	98.0	0.05	**277**	96.3	0.56	**68**	87.8	0.64	**70**

μ and σ^2 refer, respectively, to the mean and variance of the OA obtained over the 100 superpixels used. ET is presented in seconds.

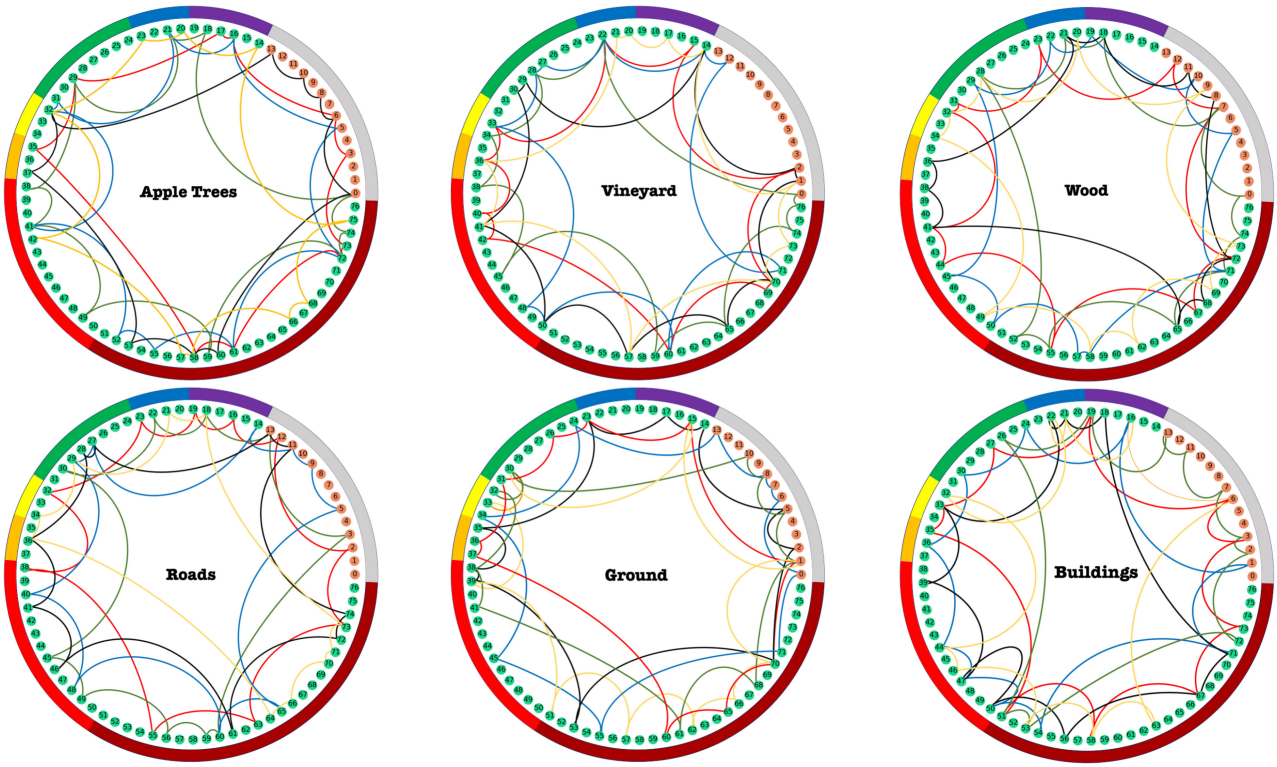


Fig. 13. Chord diagrams of selected attributes for different classes of the Trento dataset.

attributes, such as green—hyperspectral, and brown—LiDAR, while the edges illustrate the attributes that were selected by the GKMI method (color of the connections represent different homogeneous areas, i.e., superpixels). The outer circle depicts different macroscopic intervals of the spectral channels from the visible (violet 380–450 nm, blue 450–495 nm, green 495–570 nm, yellow 570–590 nm, orange 590–620 nm, and red 620–750 nm, according to the visible wavelength color representations) to the near-infrared (dark red 750–1300 nm) range with respect to the attribute numbers. The gray color represents the LiDAR attributes. Hence, the chord diagrams show that even for the same class, relevant attributes can vary and can be grouped differently. It means that if the various image parts represent the same class, they still might be observed under different technical or environmental conditions. Therefore, it is crucial to select the relevant attributes for separate zones of an image in order to reflect their particularity. The aforementioned results show the flexibility and adaptivity of the proposed information selection scheme.

5) Selected Attributes: Additionally, in order to further investigate the effectiveness of the proposed approach, we analyzed the attribute selection method with datasets that includes corrupted attributes. Accordingly, to each dataset, we added a various number of corrupted attributes, which were randomly generated by Gaussian noise with different mean $\mu = [0.1, \ldots, 1]$ and standard deviation $\sigma = [0.1, \ldots, 1]$.

Fig. 14 shows the graph of occurrences of corrupted attributes for each dataset among a different number of noisy attributes added to the original datasets. Red color refers to attributes selected by SC, while blue color demonstrates the proposed method. It can be clearly seen from the curves that there is no clear superiority of any method for Trento and Houston datasets. For a different number of noisy attributes, each of the methods shows almost equal performance, with a

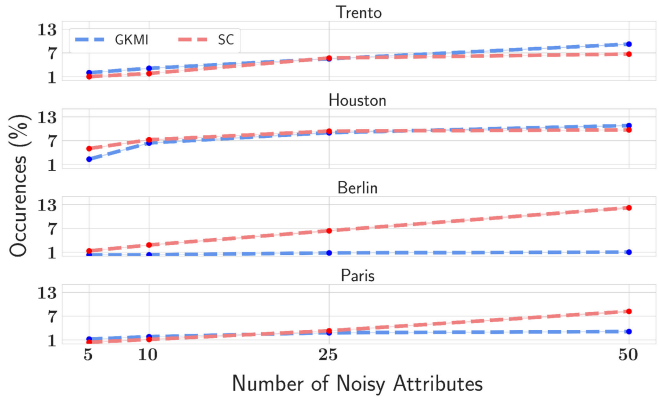


Fig. 14. Occurrences of different number of noisy attributes for SC and proposed method (GKMI).

slight advantage to one side or another. Nevertheless, for Berlin and Paris datasets, the predominance of the GKMI method becomes clearly visible. Moreover, for some parts, the percentage of selected noisy attributes using the GKMI method is several times less than using SC. Therefore, this result additionally strengthens the idea of applying two similarity metrics simultaneously.

6) Correlation Sensitivity: As an alternative metric to assess the ability of the proposed method to effective dimensionality reduction, information dependence can be taken into account. This metric is one of the commonly used criteria for feature selection, especially for hyperspectral bands that are highly correlated. Specifically, computing the Pearson correlation coefficient provides insight into the strength of a linear association between two variables. Basically, a Pearson product moment correlation attempts to draw a line of best fit through the data of two variables [82, ch. 4]. The Pearson correlation coefficient indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

To determine the strength of association based on the Pearson correlation coefficient, it is possible to rely on the amplitude of the outcome. Specifically, if the linear relationship among attributes increases, then the Pearson correlation coefficient would increase as well. Therefore, the ability of a dimensionality reduction algorithm to identify the most informative subset of features in the dataset should show up in terms of low values of Pearson correlation coefficients [82].

In our case, it is particularly important to assess the necessity of including the MI Laplacian in the dimensionality reduction process in order to improve the ability to select informative attributes in the given dataset. To demonstrate the relevance of using MI, in Table IX, we show the intercorrelation between the selected attributes using the classic SC in (5) and GKMI. Table IX reports the mean correlation and variance among all superpixels. It is evident from the results that the incorporation of MI significantly decreases the correlation of selected attributes as opposed to the SC that only utilizes the GK.

Hence, given the previous observations, employing MI in the JD procedure as in (6) appears as a key step in order to enhance the selection of relevant attributes delivered by a system based on a classic SC, especially when multimodal datasets are taken into account. Therefore, this is compliant with the results we have shown and commented on previously in this section and confirms from a statistical point of view the findings we have achieved when addressing the eigenanalysis of the selection capacity.

TABLE IX
MEAN μ AND VARIANCE σ^2 OF THE INTERCORRELATION BETWEEN THE SELECTED BANDS USING SC AND GKMI, OVER ALL SUPERPIXELS

	Berlin		Paris		Trento		Houston	
Method	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
SC	0.04	0.05	0.09	0.02	0.14	0.12	0.54	0.08
GKMI	**0.01**	0.05	**0.02**	0.02	**0.03**	0.04	**0.48**	0.01

D. Method Comparison

In order to validate the proposed attribute selection method, we compare the achieved results with nine other dimensionality reduction algorithms:

1) *one ranking approach:* FIS;
2) *two attribute extraction methods:* PCA and DBFE;
3) *three searching strategies:* FS, OBB, and GA;
4) *three graph clustering approaches:* MST clustering, DS, and CD.

The aforementioned methods were described in detail earlier in Section II. It should be emphasized that, in this work, we do not compare our method with neural-network-based approaches since they require a training set.

Tables X–XIII report the performance comparison of the GKMI method with existing methods over various multimodal datasets and using two classifiers. It is evident from the tables that graph-based approaches outperform all the classical methods for feature selection in technical literature in terms of OA, AA, and Kappa since the latter methods are not flexible enough to process the multimodal datasets. However, GKMI ensures higher accuracies over all the considered datasets with the least number of attributes since it is performed on the superpixel level. Hence, it is possible to conclude that the proposed method was finding the best descriptive attributes for each homogeneous superpixel. Moreover, it is worth noting that the two similarity measures that are employed in the GKMI scheme are apparently able to ensure a more robust definition of the connections among vertices in the graphs associated with the considered datasets. This effect allows a better characterization of the subgraphs associated with the relevant attributes. Taking into account the observations drawn previously in this work (especially when considering the parameter sensitivity analysis and the trend of the eigenvalues in Fig. 10), these results further highlight the ability of GKMI to provide robust and reliable performance in selecting the most relevant attributes under diverse analysis conditions.

TABLE X
PERFORMANCE COMPARISON AMONG DIFFERENT DIMENSIONALITY REDUCTION METHODS AND DIFFERENT CLASSIFIERS FOR THE TRENTO DATASET

Method	K	RF						SVM					
		OA (%)		Kappa	AA (%)		ET (sec)	OA (%)		Kappa	AA (%)		ET (sec)
		μ	σ^2		μ	σ^2		μ	σ^2		μ	σ^2	
PCA	26	88.1	2.1	81.2	87.8	1.9	108	88.7	1.6	88.9	88.9	1.8	**12.1**
DBFE	25	86.3	2	86.3	86.1	1.8	110.2	82.9	1.5	83.8	84.1	1.9	13.4
FIS	25	84.2	1.8	84	83.2	1.6	109.6	85.1	1.4	84.6	85.7	1.3	12.6
FS	28	84	1.2	84	83.8	1.3	110.1	82.1	1.2	81.2	82.2	1	12.8
OBB	24	90.6	0.8	89.2	90.8	0.7	124.3	90.2	0.6	87.2	90	0.7	18.8
GA	23	90.4	0.7	89.4	90.3	0.6	123.2	90.3	0.55	89.2	90.5	0.58	17.9
MST	42	95.7	0.23	94.2	95.6	0.22	**13.1**	95.6	0.07	94.1	95.8	0.05	13.5
DS	38	88.7	0.04	84.9	88.2	0.08	14.9	86.7	0.03	82.2	85.9	0.05	24.5
CD	52	96.2	0.11	94.9	96.1	0.16	16.1	96.3	0.02	95.1	96.1	0.02	14.2
GKMI	**20**	95.0	0.33	93.4	94.4	0.73	139	93.9	0.28	91.8	93.2	0.38	41
	40	**97.3**	0.14	**96.4**	**96.8**	0.28	153	**97.4**	0.29	**96.2**	**96.6**	0.56	68

TABLE XI
PERFORMANCE COMPARISON AMONG DIFFERENT DIMENSIONALITY REDUCTION METHODS AND DIFFERENT CLASSIFIERS FOR THE HOUSTON DATASET

Method	K	RF						SVM					
		OA (%)		Kappa	AA (%)		ET (sec)	OA (%)		Kappa	AA (%)		ET (sec)
		μ	σ^2		μ	σ^2		μ	σ^2		μ	σ^2	
PCA	29	79.6	1.2	79.7	79.8	1.1	188	83.2	1.1	81.0	83.6	1.3	98
DBFE	31	78.9	1.3	77.6	78.8	1.09	196	79.4	1.12	78.6	79.5	1.07	99
FIS	30	80.3	1.02	80.0	80.4	1.04	188.8	80.6	1.08	80.2	80.6	1.05	102.4
FS	30	77.5	1.3	77.1	77.5	1.21	190	77.2	1.15	76.8	77.3	1.18	100.8
OBB	26	78.5	1.11	78.4	78.7	1.09	209	77.2	1.13	75.9	77.3	1.08	116
GA	23	80.2	1.03	80.1	80.2	1.04	200.1	78.6	1.05	77.3	78.7	1.02	111
MST	40	84.9	1.68	84.0	85.6	1.71	53.5	85.7	0.72	84.7	86.3	0.65	43
DS	30	71.1	1.09	69.7	72.3	1.47	**45**	77.5	1.05	75.9	78.1	1.01	**31**
CD	64	85.0	1.98	83.9	85.6	1.79	52	85.6	0.78	84.8	86.7	0.59	47
GKMI	**20**	82.6	1.67	81.3	83.0	1.56	150	85.2	0.88	84.2	85.5	0.97	69
	60	**86.5**	1.41	**85.5**	**86.6**	1.42	161	**88.7**	0.71	**87.9**	**89.1**	0.58	88

TABLE XII
PERFORMANCE COMPARISON AMONG DIFFERENT DIMENSIONALITY REDUCTION METHODS AND DIFFERENT CLASSIFIERS FOR THE BERLIN DATASET

Method	K	RF						SVM					
		OA (%)		Kappa	AA (%)		ET (sec)	OA (%)		Kappa	AA (%)		ET (sec)
		μ	σ^2		μ	σ^2		μ	σ^2		μ	σ^2	
PCA	47	86.2	0.13	85.3	86.3	0.13	312	85.7	0.11	84.9	85.8	0.12	319
DBFE	48	86.5	0.11	86.3	86.4	0.12	313.6	85.4	0.09	85.2	85.4	0.1	320.3
FIS	47	86.6	0.11	86.5	86.8	0.12	315	86.3	0.11	85.8	86.3	0.1	323
FS	49	83.6	0.1	82.4	83.6	0.08	316.2	82.6	0.09	82.1	82.8	0.11	321
OBB	45	85.9	0.08	85.7	85.9	0.07	323	85.8	0.07	84.7	85.9	0.06	330
GA	45	86.4	0.07	86.1	86.5	0.05	320	85.7	0.04	85.2	85.8	0.05	327
MST	41	88.4	0.12	86.6	88.3	0.51	61	91.1	0.23	89.8	89.6	0.39	184
DS	68	90.1	0.12	88.6	92.8	0.08	56.8	69.6	2.51	63.6	94.6	0.02	196
CD	61	89.3	0.13	87.6	89.3	0.24	**47**	92.4	0.16	91.2	91.2	0.36	**94**
GKMI	**40**	93.5	0.22	92.6	94.9	0.17	242	94.7	0.14	93.2	93.6	0.28	243
	60	**94.2**	0.18	**93.4**	**95.5**	0.11	253	**95.5**	0.11	**94.8**	**95.2**	0.20	359

TABLE XIII
PERFORMANCE COMPARISON AMONG DIFFERENT DIMENSIONALITY REDUCTION METHODS AND DIFFERENT CLASSIFIERS FOR THE PARIS DATASET

Method	K	RF						SVM					
		OA (%)		Kappa	AA (%)		ET (sec)	OA (%)		Kappa	AA (%)		ET (sec)
		μ	σ^2		μ	σ^2		μ	σ^2		μ	σ^2	
PCA	36	88.7	0.06	87.7	88.8	0.05	331	88.4	0.08	87.9	88.5	0.07	279
DBFE	35	88.4	0.07	88.3	88.6	0.07	333	86.2	0.08	85.3	86.3	0.06	281
FIS	35	90.6	0.05	89.2	90.6	0.06	332	89.8	0.04	88.9	89.9	0.06	281
FS	37	86.3	0.06	85.6	86.3	0.05	332	86.2	0.05	85.9	86.3	0.07	283
OBB	36	87.8	0.04	87.4	87.9	0.03	344	87.5	0.03	87.1	87.6	0.02	296
GA	33	90.3	0.03	88.6	90.4	0.02	342	90.8	0.04	89.2	90.9	0.03	290
MST	41	93.8	0.06	91.8	88.6	9.83	**110**	82.4	0.15	75.3	91.0	18.03	257
DS	67	91.7	0.12	89.0	93.0	0.25	108	96.8	0.08	95.8	95.7	0.13	283
CD	83	94.0	0.02	92.0	88.4	0.98	121	82.0	0.11	74.7	91.1	0.13	443
GKMI	**30**	94.5	0.09	92.7	94.6	0.36	282	96.4	0.17	95.2	92.7	1.43	**248**
	60	**95.9**	0.07	**94.7**	**95.7**	0.22	299	**98.1**	0.03	**97.5**	**96.8**	0.46	336

V. CONCLUSION

A new unsupervised attribute selection method based on two different similarity measures has been proposed for multimodal remote sensing data. The main merits of the method are as follows.

1) *Unsupervision:* The method is application independent; therefore, it is implemented without any prior information about class labels.
2) *Flexibility:* It can be applied to datasets obtained from various sensors with different characteristics.
3) *Accuracy:* It employs two similarities that account for global and local particularities of the original dataset, which, in turn, allows selecting the most relevant attributes.
4) *Versatility:* The method is performed on a superpixel level; therefore, it selects the best descriptive attributes for each homogeneous superpixel.
5) *Interpretability:* The method retains the advantages of both attribute extraction and selection methods (preserves the physical meaning of the data and increases the separability).

The experimental results obtained from several multimodal datasets consistently demonstrated the effectiveness and robustness of the proposed method for the processing of the multimodal remote sensing datasets.

This article introduces the GKMI attribute selection method with all its crucial steps and relevant novelties. Future work directions will be focused on adding the automatic selection of the number of attributes for each superpixel, so that the multimodal data analysis can be adapted to the different conditions of the records that can be acquired on large-scale scenarios, and on developing an adaptive classifier that can deal with superpixels of heterogeneous sizes and attributes.

ACKNOWLEDGMENT

The authors would like to thank Dr. Pedram Ghamisi (Helmholtz-Zentrum Dresden-Rossendorf, Germany) for providing the Trento dataset.

REFERENCES

[1] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.
[2] M. D. Mura, S. Prasad, F. Pacifici, P. Gamba, J. Chanussot, and J. A. Benediktsson, "Challenges and opportunities of multimodality and data fusion in remote sensing," *Proc. IEEE*, vol. 103, no. 9, pp. 1585–1601, Sep. 2015.
[3] N. Longbotham *et al.*, "Multi-modal change detection, application to the detection of flooded areas: Outcome of the 2009-2010 data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 331–342, Feb. 2012.
[4] S. Chlaily, P. Amblard, O. Michel, and C. Jutten, "Impact of noise correlation on multimodality," in *Proc. 24th Eur. Signal Process. Conf.*, Aug. 2016, pp. 195–199.
[5] S. B. Serpico, M. D'Inca, F. Melgani, and G. Moser, "Comparison of feature reduction techniques for classification of hyperspectral remote sensing data," *Proc. SPIE*, vol. 4885, pp. 347–358, 2003.
[6] S. Georganos *et al.*, "Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application," *GISci. Remote Sens.*, vol. 55, no. 2, pp. 221–242, 2018.
[7] J. Bioucas-Dias *et al.*, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, Apr. 2012.
[8] C. Bichot and P. Siarry, *Graph Partitioning*. Hoboken, NJ, USA: Wiley/ISTE, 2013.
[9] U. Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, pp. 395–416, Dec. 2007.
[10] L. Gulikers, M. Lelarge, and L. Massoulié, "A spectral method for community detection in moderately sparse degree-corrected stochastic block models," *Adv. Appl. Probab.*, vol. 49, no. 3, pp. 686–721, 2017.
[11] H. T. Ali and R. Couillet, "Improved spectral community detection in large heterogeneous networks," *J. Mach. Learn. Res.*, vol. 18, no. 225, pp. 1–49, 2018.
[12] L. Dall'Amico, R. Couillet, and N. Tremblay, "Optimal laplacian regularization for sparse spectral community detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 3237–3241.
[13] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A review of unsupervised feature selection methods," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 907–948, 2020.
[14] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, 2014.
[15] M. A. Hossain, M. Pickering, and X. Jia, "Unsupervised feature extraction based on a mutual information measure for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2011, pp. 1720–1723.
[16] J. Jiang, J. Ma, C. Chen, Z. Wang, Z. Cai, and L. Wang, "SuperPCA: A superpixelwise PCA approach for unsupervised feature extraction of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4581–4593, Aug. 2018.

[17] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Orlando, FL, USA: Academic, 2008.

[18] J. Feng, L. Jiao, F. Liu, T. Sun, and X. Zhang, "Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images," *Pattern Recognit.*, vol. 51, pp. 295–309, 2016.

[19] Y. Zhou, R. Zhang, S. Wang, and F. Wang, "Feature selection method based on high-resolution remote sensing images and the effect of sensitive features on classification accuracy," *Sensors*, vol. 18, 2018, Art. no. 2013.

[20] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," in *Proc. 27th Conf. Uncertainty Artif. Intell.*, 2012, pp. 266–273.

[21] S. Sivakumar and C. Chandrasekar, "Feature selection using genetic algorithm with mutual information," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 2871–2874, 2014.

[22] P. Somol, P. Pudil, F. J. Ferri, and J. Kittler, "Fast branch & bound algorithm in feature selection," in *Proc. 4th World Multiconf. Syst., Cybern., Inform.*, Orlando, FL, USA, 2000, vol. 7, pp. 646–651.

[23] M. Fauvel, C. Dechesne, A. Zullo, and F. Ferraty, "Fast forward feature selection of hyperspectral images for classification with Gaussian mixture models," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2824–2831, Jun. 2015.

[24] M. Ahmad, D. Ulhaq, Q. Mushtaq, and M. Sohaib, "A new statistical approach for band clustering and band selection using K-means clustering," *Int. J. Eng. Technol.*, vol. 3, pp. 606–614, Dec. 2011.

[25] W. Sun and Q. Du, "Hyperspectral band selection: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 118–139, Jun. 2019.

[26] J. Feng, L. Jiao, T. Sun, H. Liu, and X. Zhang, "Multiple kernel learning based on discriminative kernel clustering for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 11, pp. 6516–6530, Nov. 2016.

[27] S. Schaeffer, "Graph clustering," *Comput. Sci. Rev.*, vol. 1, pp. 27–64, 2007.

[28] M. T. Altuncu, E. Mayer, S. N. Yaliraki, and M. Barahona, "From free text to clusters of content in health records: An unsupervised graph partitioning approach," *Appl. Netw. Sci.*, vol. 4, 2019, Art. no. 2.

[29] M. Schaub, J.-C. Delvenne, R. Lambiotte, and M. Barahona, "Multiscale dynamical embeddings of complex networks," *Phys. Rev. E*, vol. 99, 2019, Art. no. 062308.

[30] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond euclidean data," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, Jul. 2017.

[31] T. Berry and T. Sauer, "Consistent manifold representation for topological data analysis," *Found. Data Sci.*, vol. 1, no. 1, pp. 1–38, 2019.

[32] M. Carreira-Perpiñán and R. Zemel, "Proximity graphs for clustering and manifold learning," in *Proc. 17th Int. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2004, pp. 225–232.

[33] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[34] O. Grygorash, Y. Zhou, and Z. Jorgensen, "Minimum spanning tree based clustering algorithms," in *Proc. 18th IEEE Int. Conf. Tools Artif. Intell.*, 2006, pp. 73–81.

[35] M. Beguerisse-Díaz, B. Vangelov, and M. Barahona, "Finding role communities in directed networks using role-based similarity, Markov stability and the relaxed minimum spanning tree," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Austin, TX, USA, 2013, pp. 937–940.

[36] R. Liu, S. Feng, R. Shi, and W. Guo, "Weighted graph clustering for community detection of large social networks," *Procedia Comput. Sci.*, vol. 31, pp. 85–94, 2014.

[37] M. Pavan and M. Pelillo, "Dominant sets and pairwise clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 167–172, Jan. 2007.

[38] R. Lambiotte, J. Delvenne, and M. Barahona, "Random walks, Markov processes and the multiscale modular organization of complex networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 1, no. 2, pp. 76–90, Jul.–Dec. 2014.

[39] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, "Stability of graph communities across time scales," *Proc. Nat. Acad. Sci.*, vol. 107, no. 29, pp. 12755–12760, 2010.

[40] R. Tripodi, S. Vascon, and M. Pelillo, "Context aware nonnegative matrix factorization clustering," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2016, pp. 1719–1724.

[41] S. Vascon, M. Cristani, M. Pelillo, and V. Murino, "Using dominant sets for k-NN prototype selection," in *Proc. Int. Conf. Image Anal. Process.*, 2013, pp. 131–140.

[42] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[43] X. Xu, H. Gu, Y. Wang, J. Wang, and P. Qin, "Autoencoder based feature selection method for classification of anticancer drug response," *Front. Genetics*, vol. 10, 2019, Art. no. 233.

[44] D. Tomar, Y. Prasad, M. K. Thakur, and K. K. Biswas, "Feature selection using autoencoders," in *Proc. Int. Conf. Mach. Learn. Data Sci.*, Dec. 2017, pp. 56–60.

[45] R. Couillet and M. McKay, "Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators," *J. Multivariate Anal.*, vol. 131, pp. 99–120, 2014.

[46] J. Hu, D. Hong, Y. Wang, and X. Zhu, "A comparative review of manifold learning techniques for hyperspectral and polarimetric SAR image fusion," *Remote Sens.*, vol. 11, no. 6, pp. 1–28, 2019.

[47] D. Tuia, M. Volpi, M. Trolliet, and G. Camps-Valls, "Semisupervised manifold alignment of multimodal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7708–7720, Dec. 2014.

[48] D. Hong, J. Kang, N. Yokoya, and J. Chanussot, "Graph-induced aligned learning on subspaces for hyperspectral and multispectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4407–4418, May 2021.

[49] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[50] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[51] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal GANs: Toward crossmodal hyperspectral-multispectral image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5103–5113, Jun. 2021.

[52] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-modalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 12–23, 2020.

[53] G. Iyer, J. Chanussot, and A. L. Bertozzi, "A graph-based approach for data fusion and segmentation of multimodal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4419–4429, May 2021.

[54] J. Xia, W. Liao, and P. Du, "Hyperspectral and Lidar classification with semisupervised graph fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 666–670, Apr. 2020.

[55] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, "Learnable manifold alignment (LEMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 147, pp. 193–205, 2019.

[56] J. Dy and C. Brodley, "Feature selection for unsupervised learning," *J. Mach. Learn. Res.*, vol. 5, pp. 845–889, Aug. 2004.

[57] S. Doan and S. Horiguchi, "An efficient feature selection using multi-criteria in text categorization for Naïve Bayes classifier," *WSEAS Trans. Inf. Sci. Appl.*, vol. 2, no. 2, 2005, Art. no. 34.

[58] L. Rokach, B. Chizi, and O. Maimon, "Feature selection by combining multiple methods," in *Advances in Web Intelligence and Data Mining*, vol. 23. New York, NY, USA: Springer, 2006, pp. 295–304.

[59] J. Shi and J. Malik, "Normalized cuts and image segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 1997, pp. 731–737.

[60] G. Strang, *Linear Algebra and its Applications*. Belmont, CA, USA: Thomson, Brooks/Cole, 2006.

[61] P. Ablin, J. Cardoso, and A. Gramfort, "Beyond Pham's algorithm for joint diagonalization," 2018, *arXiv:1811.11433*.

[62] Dinh-Tuan Pham and J. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.

[63] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2001, pp. 849–856.

[64] A. Plaza and C.-I. Chang, *High Performance Computing in Remote Sensing*. New York, NY, USA: CRC Press, 2007.

[65] Y. Liu, Q. Ren, J. Geng, M. Ding, and J. Li, "Efficient patch-wise semantic segmentation for large-scale remote sensing images," *Sensors (Switzerland)*, vol. 18, no. 10, pp. 1–16, 2018.

[66] S. Beucher, "The watershed transformation applied to image segmentation," in *Proc. 10th Pfefferkorn Conf. Signal Image Process. Microsc. Microanal.*, 1992, pp. 299–314.

[67] P. Neubert and P. Protzel, "Compact watershed and preemptive SLIC: On improving trade-offs of superpixel segmentation algorithms," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 996–1001.

[68] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[69] A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Evaluating the performance of a random forest kernel for land cover classification," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 575.
[70] J. Xia, N. Falco, J. Benediktsson, P. Du, and J. Chanussot, "Hyperspectral image classification with rotation random forest via," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1601–1609, Apr. 2017.
[71] S. R. Gunn, "Support vector machines for classification and regression," 1998.
[72] A. Liaw and M. Wiener, "Classification and regression by random forest," *Forest*, vol. 23, pp. 18–23, Nov. 2001.
[73] P. Bharatkar and R. Patel, "Approach to accuracy assessment tor RS image classification techniques," *Int. J. Sci. Eng. Res.*, vol. 4, no. 12, pp. 79–86, 2013.
[74] R. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.
[75] U. Kandaswamy, D. A. Adjeroh, and M. C. Lee, "Efficient texture analysis of SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 9, pp. 2075–2083, Sep. 2005.
[76] F. Albregtsen, "Statistical texture measures computed from gray level coocurrence," *Boundary 2*, vol. 3, no. 1, p. 45, 1974.
[77] *2017 IEEE GRSS data fusion contest*, 2017. [Online]. Available: http://www.grss-ieee.org/community/technical-committees/data-fusion/2017-ieee-grss-data-fusion-contest-2/.
[78] *2013 IEEE GRSS data fusion contest*, 2013. [Online]. Available: http://www.grss-ieee.org/community/technical-committees/data-fusion/2013-ieee-grss-data-fusion-contest/.
[79] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*. New York, NY, USA: Wiley, 2009.
[80] S. Prasad, L. Bruce, and J. Chanussot, *Optical Remote Sensing, Augmented Vision and Reality*. Berlin, Germany: Springer, 2011.
[81] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "A spatial-spectral kernel-based approach for the classification of remote-sensing images," *Pattern Recognit.*, vol. 45, no. 1, pp. 381–392, 2012.
[82] R. Rousseau, L. Egghe, and R. Guns, Eds., *Becoming Metric-Wise* (Chandos Information Professional Series). Amsterdam, The Netherlands: Elsevier, 2018.

Eduard Khachatrian (Student Member, IEEE) received a double M.Sc. degrees in polar and marine sciences from the Faculty of Mathematics, Informatics, and Natural Sciences, University of Hamburg, Hamburg, Germany, and the Institute of Earth Sciences, Saint Petersburg State University, Saint Petersburg, Russia, in 2017. He is currently working toward the Ph.D. degree with the Center of Integrated Remote Sensing and Forecasting for Arctic Operations, University of Tromsø—The Arctic University of Norway, Tromsø, Norway.

From 2017 to 2018, he was a Junior Scientist with the Nansen International Environmental and Remote Sensing Centre, Bergen, Norway. His research interests include multimodal data analysis, image processing, and remote sensing of polar areas.

Saloua Chlaily (Member IEEE) received the M.Sc. degree in electronics engineering from the École Nationale Supérieure d'Électrotechnique, d'Électronique, d'Informatique, d'Hydraulique et des Télécommunications, Toulouse, France, and the M.Sc. degree in electrical engineering from the Hassania School of Public Works, Casablanca, Morocco, both in 2013, and the Ph.D. degree in signal, image, speech, and telecommunication from Grenoble Alpes University, Grenoble, France, in 2018.

She is currently a Researcher with University of Tromsø—The Arctic University of Norway (UiT). She conducts her research with the Center of Integrated Remote Sensing and Forecasting for Arctic Operations, UiT. Her research interests include multimodal data analysis, signal processing, and image processing.

Dr. Chlaily is a Reviewer for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

Torbjørn Eltoft (Member, IEEE) received the M.Sc. and Ph.D. degrees from the University of Tromsø, Tromsø, Norway, in 1981 and 1984, respectively.

In 1988, he joined the Department of Physics and Technology, University of Tromsø—The Arctic University of Norway (UiT), Tromsø, where he is currently a Professor and the Director of the Centre for Integrated Remote Sensing and Forecasting for Arctic Operations, a center for research-based Innovation awarded by the Norwegian Research Council in 2014, whose objective is to develop knowledge and remote sensing technology for arctic applications. From 2013 to 2015, he was the Head of the Department of Physics and Technology, UiT. He was an International Researcher with the University of California, Irvine, CA, USA, from 1992 to 1993 and then from 1997 to 1998, and with the University of California, San Diego, CA, from 2004 to 2005. He has a significant publication record in the areas of signal processing and remote sensing. His research interests include multidimensional signal and image analysis, statistical modeling, neural networks, and machine learning, with emphasis on applications in multichannel synthetic aperture radar remote sensing and multisensor remote sensing for the Arctic.

Dr. Eltoft was an Associate Editor for *Pattern Recognition* from 2005 to 2011 and a Guest Editor for *Remote Sensing*'s Special Issue for the PolInSAR 2017 Conference. He was the co-recipient of the Outstanding Paper Award in Neural Networks awarded by the IEEE Neural Networks Council in 2000, Honorable Mention for the 2003 *Pattern Recognition* Journal Best Paper Award, and the 2017 UiT Award for Research and Development from the UiT.

Andrea Marinoni (Senior Member, IEEE) received the B.S., M.Sc. (*cum laude*), and Ph.D. degrees in electronic engineering from the University of Pavia, Pavia, Italy, in 2005, 2007 and 2011, respectively.

He is currently an Associate Professor with the Earth Observation Group, Centre for Integrated Remote Sensing and Forecasting for Arctic Operations, Department of Physics and Technology, University of Tromsø—The Arctic University of Norway, Tromsø, Norway, and a Visiting Academic Fellow with the Department of Engineering, University of Cambridge, Cambridge, U.K. From 2013 to 2018, he was a Research Fellow with the Telecommunications and Remote Sensing Laboratory, Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy. In 2009, he has been a Visiting Researcher with the Communications Systems Laboratory, Department of Electrical Engineering, University of California, Los Angeles, CA, USA. In 2011, he was the recipient of the two-year "Applied research grant," sponsored by the Region of Lombardy, Italy, and STMicroelectronics N.V. In 2017, he was the recipient of the INROAD grant, sponsored by the University of Pavia and Fondazione Cariplo, Italy, for supporting excellence in design of European Research Council proposal. In 2018, he was the recipient of the "Progetto professionalità Ivano Becchi" grant funded by the Fondazione Banco del Monte di Lombardia, Italy, and sponsored by the University of Pavia and the NASA Jet Propulsion Laboratory, Pasadena, CA, for supporting the development of advanced methods of air pollution analysis by remote sensing data investigation. He was the recipient of Asgard Research Program and Asgard Recherche+ Program grants funded by the Institut Français de Norvége, Oslo, Norway, in 2019 and 2020, respectively, for supporting the development of scientific collaborations between French and Norwegian research institutes. From 2015 to 2017, he was a Visiting Researcher at the Earth and Planetary Image Facility, Ben Gurion University of the Negev, Be'er Sheva, Israel; the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China; the School of Computer Science, Fudan University, Shanghai, China; the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China; and the Instituto de Telecomunicações, Instituto Superior Tecnico, Universidade de Lisboa, Lisbon, Portugal. In 2020 and 2021, he was a Visiting Professor with the Department of Electrical, Computer and Biomedical Engineering, University of Pavia. His main research interests include efficient information extraction from multimodal remote sensing, nonlinear signal processing applied to large-scale heterogeneous records, Earth observation interpretation and Big Data mining, and analysis and management for human–environment interaction assessment.

Dr. Marinoni is the Founder and Current Chair of the IEEE Geoscience and Remote Sensing Society (GRSS) Norway Chapter. He is also an Ambassador for IEEE Region 8 Humanitarian activities, and a research contact point for the Norwegian Artificial Intelligence Research Consortium. He serves as a Topical Associate Editor of machine learning for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He was a Guest Editor of three special issues on Multimodal Remote Sensing and Sustainable Development for IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING. He is the Leader of the GR4S Committee of the IEEE GRSS, coordinating the organization of schools and workshops sponsored by the IEEE GRSS worldwide.