Feature-Free Explainable Data Mining in SAR Images Using Latent Dirichlet Allocation

Chandrabali Karmakar, Corneliu Octavian Dumitru, Gottfried Schwarz, and Mihai Datcu, *Fellow, IEEE*

***Abstract*—In this article, we propose a promising approach for the application-oriented content classification of spaceborne radar imagery that presents an interesting alternative to popular current machine learning algorithms. In the following, we consider the problem of unsupervised feature-free satellite image classification with already known classes as an explainable data mining problem for regions with no prior information. Three important issues are addressed here: explainability, feature independence, and unsupervision. There is an increasing demand toward explainable machine learning models as they strive to meet the "right to explanation." The importance of feature-free classification stems from the problem that different classification outcomes are obtained from using different features and the complexity of computing sophisticated image primitive features. Developing unsupervised discovery techniques helps overcome the limitations in object discovery due to the lack of labeled data and the dependence on features. In this article, we demonstrate the applicability of a latent Dirichlet allocation (LDA) model, one of the most established unsupervised probabilistic methods, in discovering the latent structure of synthetic aperture radar data. The idea is to use LDA as an explainable data mining tool to discover scientifically explainable semantic relations. The suitability of the approach as an explainable model is discussed and interpretable topic representation maps are produced, which practically demonstrate the idea of "interpretability" in an explainable machine learning paradigm. LDA discovers the latent structures in the data as a set of topics. We create the interpretable visualizations of the data utilizing these topics and compute the topic distributions for each land-cover class. Our results show that each class has a distinct topic distribution that represents that particular class. Then these classes can be grouped based on their similarity of topic composition. Both the topic composition and grouping are explainable by domain experts.**

***Index Terms*—Bag of words technique, discovery, explainable machine learning, interpretability, latent Dirichlet allocation (LDA), synthetic aperture radar (SAR), unsupervised image classification.**

I. Introduction

IN SPITE of the availability of immense amounts of earth observation data from various sensors and many artificial intelligence algorithms, the explainable unsupervised classification of earth observation images is still an unexplored area. Still, the question remains as "How to make a classification with little or no labeled data following a glass-box approach?" Here, we have two problems: first, the unavailability of labeled datasets; and second, explainability. Most machine learning models in use today are supervised and trained on some labeled data. In practice, first, obtaining labeled data are very expensive as it needs human expertise. Not only that, any inaccuracy in the labeled data leads to inaccuracies in the model. Naturally, it is beneficial to be able to explore and make use of the huge amounts of unlabeled being data available. However, the task of image annotation or classification becomes challenging with a limited amount of labeled data. In [25], most of the work refers to the labeling of the data and matching the semantic classes based on the empirical probability matrix. Unsupervised classification is traditionally made with clustering techniques; however, there are some limitations of clustering approaches. Hierarchical clustering algorithms (such as the k means and agglomerative complete-link algorithms) depend on the selected similarity measure. Not only that, clustering in the high-dimensional data is very time consuming. As of today, there have already been some efforts in this direction.

The next question to answer is "Explainability." Machine learning techniques have shown undeniable success in various disciplines. However, there has been an emerging demand to understand how a model functions and an explanation of its output. Deep learning models have shown tremendous success in various areas from image understanding and natural language processing to speech recognition. However, most of these approaches do not offer any explainability of the method. In many applications, it may be unacceptable to trust the decisions of a black-box system. For instance, in societal contexts, the reasons for a decision are important. Typical examples are (semi) automatic loan applications, employment decisions, or risk assessments for insurance applicants, where it must be clear to the user why a model gives a particular prediction and how such predictions can affect him/her. In this context, and also due to regulatory reasons, one aim is that decisions based on the machine learning models shall be fair and ethical. The importance to give reasons for the decisions of a machine learning algorithm is also high for medical applications, where motivation is the question of trust in decisions so that patients can rely on a decision having been made. All this is supported by the General Data Protection Regulation of the European Union, which contains new rules regarding the use of personal information.

One component of these rules can be summed up by the term "right to explanation" [1]. According to Adadi and Berrada [4],

Manuscript received July 14, 2020; revised October 18, 2020 and November 13, 2020; accepted November 14, 2020. Date of publication November 18, 2020; date of current version January 6, 2021. This work was supported in part by the Helmholtz Automated Scientific Discovery Project and in part by the H2020 ExtremeEarth Project under Grant 825258. *(Corresponding author: Corneliu Octavian Dumitru.)*

The authors are with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany (e-mail: chandrabali.karmakar@dlr.de; corneliu.dumitru@dlr.de; gootfried.schwarz@dlr.de; mihai.datcu@dlr.de).

Digital Object Identifier 10.1109/JSTARS.2020.3039012

there are mainly four reasons to ask for explanations: to get clear reasons for the decisions; to have greater control over the model; to improve models; to discover new knowledge. Researchers in the field [5] claim explainability as a prerequisite to ensure the scientific value of the outcomes. Another important argument about having an explainable machine learning model is that it satisfies the "right to explanation" [1].

Recently, there have been efforts toward explaining black-box AI models. Schlegel *et al*. [19] present a methodology to test and evaluate various XAI methods on time series. Hohman *et al*. [20] present a survey of the role of visual analytics in deep learning research to understand how models work. Xi and Panoutsos [21] propose a convolutional neural network learning structure with added interpretability-oriented layers, in the form of fuzzy logic-based rules. The previously mentioned researches focus on understanding deep learning models, while Wick *et al.* [22] propose a "cyclic boosting" machine learning algorithm, which allows us to efficiently perform accurate regression and classification tasks, and also allows detailed understanding of each individual prediction made by the model.

In our case, we address the problem of explainable unsupervised image classification with explainable data mining based on the latent Dirichlet allocation (LDA). The purpose of such research lies in proposing a discovery method, which follows the principles of explainable machine learning and is seamless. Our approach aims to "discover" latent structures in the dataset using LDA and explain the final outcomes in terms of the physical world.

The reasons behind using LDA are as follows.

1) It is completely unsupervised.
2) It gives interpretable intermediate results and explainable outcomes.
3) Having LDA as the learning model in our method helps us satisfy most conditions of explainable machine learning.

In practice, it is unrealistic to expect that a single method can satisfy all aspects of explainable machine learning as listed by Roscher *et al*. [5]. However, we show that our LDA-based approach supports the three main aspects of explainable machine learning, i.e., transparency, interpretability, and explainability, and can be considered as a highly explainable approach. LDA originated as a topic model and has found wide acceptance in remote sensing image processing. Additionally, an adaptation of probabilistic latent semantic analysis was proposed to fuse the synthetic aperture radar (SAR) and multispectral (optical) imaging data for unsupervised land-cover categorization tasks [6].

LDA has been used on panchromatic Quickbird images to annotate large satellite images using semantic concepts, where examples of each semantic concept are given by the user [7]. LDA has also been used for measuring changes in multispectral image time series [8]. LDA was also applied to high-level scene understanding and content extraction [9], which aims to discover latent semantic classes containing pairs of objects characterized by a certain spatial positioning. Tănase *et al*. [10] used LDA to discover semantic relationships in PolSAR images. Bratasanu *et al.* [11] use LDA to map heterogeneous pixels with similar intermediate-level semantics called topics into separate classes. Here, we go further in the use of LDA. We use LDA as an explainable machine learning model to produce interpretable topic representation maps, explainable semantic relations, and class distinctions based on semantic class-topic relations.

A. Characteristics of Our Approach (Already Applied in a Sea-Ice Case Study)

The general workflow of our method is presented in Fig. 1. We used the annotated sea-ice classes from an active learning research project [15] as a case study to retrieve semantic relations with intermediate-level semantics (topics) using LDA. The topics retrieved from LDA are subjected to some data analysis to obtain the semantic relations. Then, a domain expert can explain these semantic relations of the sea-ice classes. At this point, we introduced the following unique characteristics of our approach from a methodological point of view.

1) *Feature-independence and topics as intermediate-level semantics*: As there was no prior information, we used a seamless data mining approach using LDA to discover the semantic relations in the land-cover classes retrieved by another research project [15]. The results allow for interpretability and explainability due to their seamlessness. For instance, we proposed several methods to generate topic representation maps from seamless features, which are fit for visual interpretation when compared with a corresponding product quick-look image, without needing any reference classification map. We also proposed methods to derive explainable semantic relations between intermediate-level semantics from LDA and the reference classes, and explainable similarity information among these classes based on their semantic relations.
2) *Explainable semantic relations*: The semantics from our unsupervised model are the intermediate-level semantics called topics. The semantic relations between such topics and the supervised classification results from [15] provide a mapping from the knowledge obtained from discovery to that from a supervised classification. This can, possibly, provide more granularity in the knowledge. The semantic relations are further used to demonstrate the composition of the land-cover classes. The difference between each pair of classes is also provided in terms of semantic relations, i.e., the presence, absence, and abundance of each topic in quantitative terms.
3) *Interpretable topic representation*: Topic representation maps are the initial visual information from the latent topics discovered by LDA. This provides an educated guess about the study area without requiring any reference data. We chose the maximum probable topic for each word in the bag of words framework to create this topic representation map; consequently, the representation is, in turn, ranked by the trust levels of the topic probabilities. From the point of view of explainable machine learning, this visualization serves as a tool for interpretation, possibly giving more granular details.

   Another important usefulness of topic representations is that it is independent of any reference data. The topic representation map can be compared with the quick-look image of the product to make further interpretations.

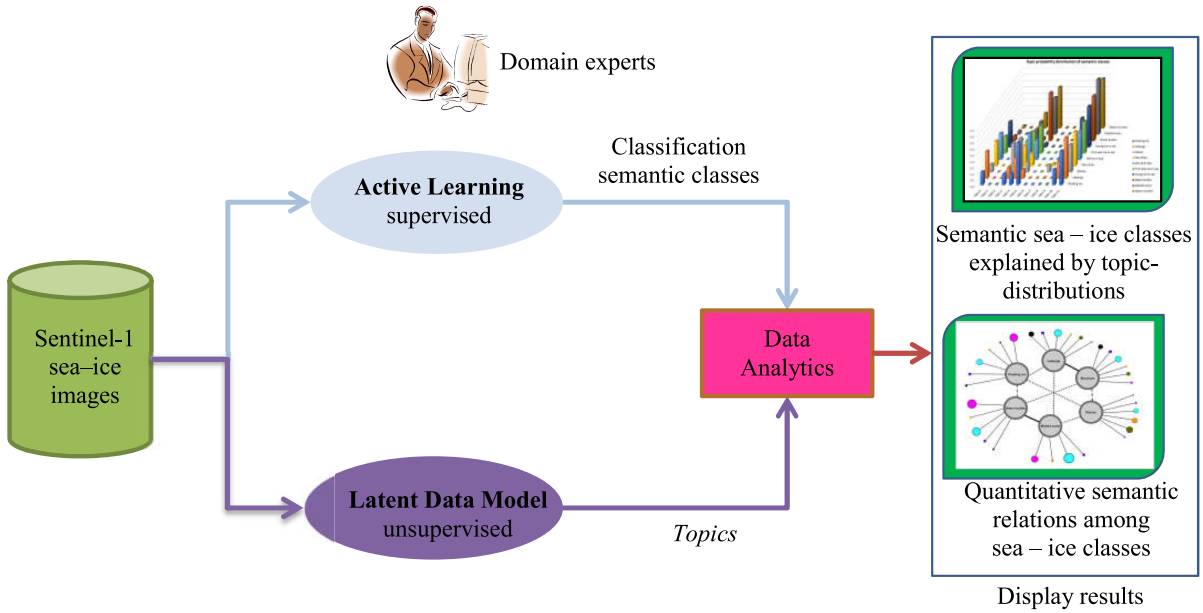


Fig. 1. Basic LDA workflow.

4) *Explainable class distinctions as a validation tool*: We used a grouping of classes based on their semantic relations, which is explainable by a domain expert. We propose it also as a validation method, which does not need any additional verification information.

In a nutshell, the complete method is dedicated to explore unknown data with no training or verification information. The rest of the article is organized as follows. Section II presents the concept of explainable machine learning and the adherence of our approach to the principles of explainable machine learning, while Section III presents the dataset being used for our experiments and the general workflow of our method. Section IV explains how LDA was used to derive semantic relations and the explainability of the outcomes of this step. Section V discusses the interpretable topic representation map, while Section VI is dedicated to the class distinction step and the explainability of the results. The article is concluded with a discussion in Section VII.

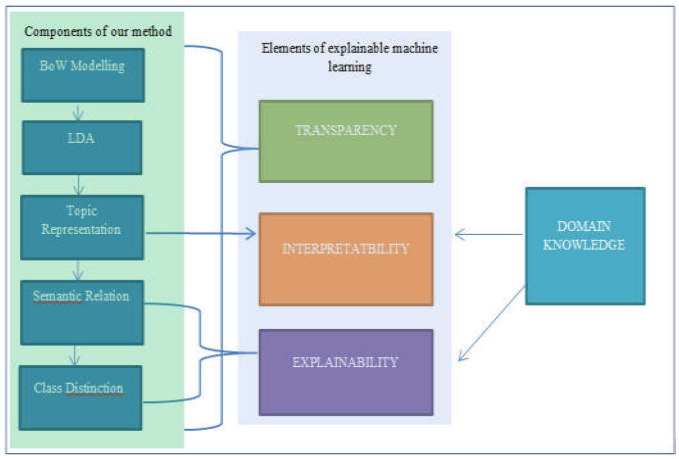


Fig. 2. LDA as the explainable data mining. The diagram shows the contribution of each component of our method to the components of explainable machine learning.

II. Explainable Machine Learning

The relevant literature on explainable machine learning lists three aspects of it: transparency, interpretability, and explainability. In Fig. 2, we describe how the components of our method satisfy these aspects and present them diagrammatically. However, it is important to mention that there are various degrees of completeness in models describing the ideas of explainable machine learning. In other words, some models are more explainable than others as they fulfill more conditions of explainability, achieved by the application of the aspects of explainable machine learning: transparency, interpretability, and explainability.

Informally, transparency is the capability of understanding the mechanism of each component of a method. Lipton [12] delineates three levels of transparency: design transparency, algorithmic transparency, and model transparency. Design transparency calls for a clear logic behind the design decisions, such as model parameters, choice of a distance metric, e.g., the Euclidian metric, and choosing between linear and nonlinear kernels. Algorithmic transparency is the capability to understand how the algorithm works from a mathematical point of view. A model is called algorithmically transparent if its input–output relations and the process can be written down as a mathematical formula [5]. On the other hand, approximations, such as early stopping and stochastic gradient descent, cause algorithmic nontransparency. Finally, the model transparency ensures the traceability of the outcomes. All these types of transparency do not depend on specific data but on the method being followed. Roscher *et al.* [5] mention [13] as a black-box approach, commenting that

it is design-nontransparent, noninterpretable, and nonexplainable as it neither offers any explanation for its design choices nor applies any domain knowledge to explain its outcomes. In contrast to this, our method is highly design-transparent, model-transparent, and algorithmically transparent to a great extent. A minor amount of nontransparency is introduced by the use of LDA as the posterior distribution is intractable, and approximations have to be used. However, the repeated application of the method has shown consistent outcomes, thereby ensuring reproducibility, overcoming the typical drawbacks introduced by LDA, and making the whole method transparent with simulatable outcomes.

The first step of our method is bag-of-words modeling. This step satisfies the three levels of transparency: design, algorithmic, and model transparency. The method is design-transparent as we have a clear reason behind choosing the number of words, i.e., the vocabulary size is algorithmically transparent as one can clearly understand how the modeling is done as described above, and is model-transparent because the results are reproducible. For the next step, i.e., the LDA modeling, we can claim that it is design-transparent, as we already learned the parameters of LDA from the data itself, and we set the number of topics empirically based on the assumption that the number of topics should not be less than the number of classes found by the reference research in [15]. LDA is not completely algorithmically transparent as the posterior distribution of LDA is intractable; however, the model transparency is achieved as we get similar results with repeated runs. The next step of our method is the topic representation. The idea of design transparency does not apply here; however, this step is model and algorithmically transparent. The same holds for the next two steps, interoperability and explainability.

The second aspect of explainable machine learning is interpretability, which is nothing but making sense of intermediate outcomes, e.g., from the latent layers of the model in combination with the help of domain knowledge. Currently, several interpretation tools are used by researchers. Among them, visualization tools are the most commonly used ones [5]. The model used in [14] is mentioned by Roscher *et al.* [5] as an interpretable model. As an alternative, Ghosal *et al.* [14] produce the maps of disease symptoms for plant stress phenotyping to be compared with the manual maps produced by experts, thereby interpreting the intermediate outcomes. Similar to this, we produce the topic representation maps to visually interpret topics (our latent layer retrieved from LDA) and by comparing the topic representation maps with the classification maps produced by supervised classification, as described in [15], we also interpretable outcomes from the third step, i.e., topic representation. Technically, topics are retrieved in the second step (LDA) and they become interpretable with the help of the visualizations developed in the topic representation step. We can compare these topic representation maps with the classification maps to make interpretations about the topics, thus providing explainable class distinctions.

Next comes the third aspect, explainability. Montavon *et al.* [3] define explainability as a decision produced with the help of collections of elements in the interpretable domain, together with the domain knowledge.

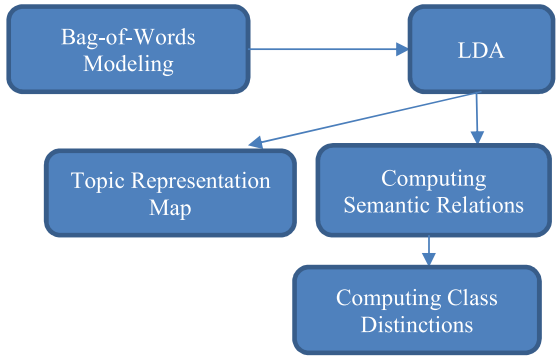


Fig. 3. Detailed workflow.

Here the topics are latent layer variables holding the intermediate-level representations of the data that fall into the interpretable domain. We then utilize these topics to create explainable outcomes: explainable semantic relations and explainable class distinctions based on these semantic relations. The concepts are presented in greater detail in Sections IV and V. The final outputs are obtained from the semantic relations and the class distinction steps. Finally, our method produces scientifically consistent and explainable outcomes about the compositions of classes.

III. Dataset and General Workflow

The proposed methodology was validated using Sentinel-1 satellite data acquired around the North Pole. The images are accessible for download from the ESA's Copernicus Hub. Sentinel-1 is a *C*-band SAR instrument and its characteristics are described in [15] in detail.

From the available Sentinel-1 data products, we selected for demonstration, based on our previous experience, level-1 ground range detected data with high resolution taken in the interferometric wide swath mode. The products are geocoded with a resolution of 20 × 22 m (range × azimuth) and a pixel spacing of 10 × 10 m. For these products, the images are provided in dual polarization (for our polar areas, HH and HV) and with an incidence angle of about 45°. The data are amplitude data.

In this article, we propose a promising approach for the application-oriented content classification of spaceborne radar imagery that presents an interesting alternative to popular current machine learning algorithms. In the following, we chose these images because they are free to access, offer frequent acquisitions over the investigated areas, and are not dependent on the prevailing weather conditions.

Here, both polarizations (HH and HV) are combined based on [17].

The proposed method contains five main steps as demonstrated in Fig. 3.

A. Experimental Setting

Macropatches with a size 256 × 256 pixels were extracted from the image dataset consisting of three scenes; with a size of each scene being 25 673 × 16 641 pixels, a total of 19 500

patches were extracted. The reason for doing this was to match the setting of the reference classes from [15]. Furthermore, from these 19 500 macropatches, the micropatches of size 4 $\times$ 4 were extracted by dense sampling, i.e., 4096 micropatches from each macropatch (of 256 $\times$ 256 pixels). The size of the macropatches was intensively validated for SAR data as follows: for TerraSAR-X data, as described in [23]; and for Sentinel-1 data, as described in [24]. In addition to this, an analysis of the size of the micropatches was made in [18].

The pixel values in these micropatches were taken as the local descriptors. A k means clustering was applied to these local descriptors to obtain a 50-word dictionary [8]. After the word assignment, we computed the histograms of words for each macropatch, thereby generating the bag of words representation of the dataset. Considering the macropatches as visual documents, the micropatches as visual words, and the set of 19 500 visual documents as a single visual corpus, we subjected this visual corpus to LDA with the number of topics set to 12 to roughly match the assumed number of semantic classes. A gensim implementation of LDA written in python programming language was used. The model calculates the sparsity parameters from the data. This only states the experimental setting for generating a bag of words model and applying LDA; the detailed procedures of computing semantic relations, interpretable visualizations, class distinctions (for validation), and the rationales behind the design choices are presented in detail in the respective sections.

The pseudocode listed in the following part is presented as supplementary information to help readers to follow the method. The workflow diagram in Fig. 3 shows the main steps in our approach. The first step is the bag of words modeling. This step is necessary before applying LDA to the image dataset. The next step is the LDA modeling; this step learns the parameters from the data and then models the Sentinel-1 scenes as the distributions of latent variables called "topics." The fourth step is for computing semantic relations between the reference classes and the latent variables retrieved from the LDA model. The fifth step creates a topic representation map, while the final step is to compute the class distinctions based on precomputed semantic relations. The underlying concepts are presented in the following sections.

Step 0: Inputs
-A database of images I_i ($i = 1, 2, 3, \ldots, N$).
-Size of each image (H_i, W_i) pixels.
-Class labels C_j ($j = 0, 1, 2, \ldots,$ NUM_OF_CLASSES-1) for M patches (size: P_h, P_w pixels) cut out from N images.
-Dictionary size V, Number of topics K.
-Size of micropatches (R_h, R_w)

Step 1: #This part computes the bag of words model of images
#Extract macro and micropatches
For i **in** (0, N-1):
MACROPATCHES = *extract_patches* (I_i,P_h,P_w)
For m **in** (0, *length*(MACROPATCHES)-1):
MICROPATCHES = *extract_patches*(MACROPATCHES$_m$, R_h, R_w)
#Dictionary creation
FEATURE_MATRIX = *reshape*(MICROPATCHES, (*length*(MICROPATCHES), ($R_h * R_w *$ NUM_OF_CHANNELS))
DICTIONARY, CLUSTERLABELS = k-means (FEATURE_MATRIX, VOCABULARY_SIZE = V)
DATA_ARRAY = *reshape*(CLUSTERLABELS, (P_h, P_w))
BOW = *compute_histogram*(DATA_ARRAY, NUMBER_OF_BINS = V)
Return BOW

Step 2: #This part applies LDA to the bag of words model computed above
#Learn the alpha and beta parameters
ALPHA_BETA = *learn_alpha_beta*(BOW)
WORD_TOPIC_PROBABILITIES, DOCUMENT_TOPIC_PROBABILITIES = *lda_model*(DATA_ARRAY, K)
Return (both) PROBABILITIES

Step 3: #This part computes the semantic relations
CLASS_TOPIC_PROB_MAT = *zeros*(NUM_OF_CLASSES, K)
SEMANTIC_RELATION = *zeros*(NUM_OF_CLASSES, K)
#Computation
For j **in** (0, NUM_OF_CLASSES-1):
  For k in (0, K-1):
    CLASS_TOPIC_MAT[j, k] = *count_occurences*(CLASS = j, TOPIC = k, *argmax_topic_probability*(DATA_ARRAY, WORD_TOPIC_PROBABILITIES))
    SEMANTIC_RELATION[j, k] = CLASS_TOPIC_MAT[j, k]/*number_of_micropatches*(Cj)
Return SEMANTIC_RELATIONS

Step 4:#This part draws the topic representation maps
COLORS = COLOR$_i$.($i = 0, 1, \ldots, K$-1)
draw_topic_representation(DATA_ARRAY, COLORS, WORD_TOPIC_PROBABILITIES))
#And save the map on disk

Step 5:#Class distinction
avg_kl_div(P_X, Q_X) = (*sum*($P\_X[i]\cdot\log(P\_X[i]/Q\_X[i])+$ *sum*($Q\_X[i]\cdot\log(Q\_X[i]/P\_X[i])$)/2
For i **in** (0, NUM_OF_CLASSES-1):
  For j **in** (0, NUM_OF_CLASSES-1):
    DISTANCE_MAT[i, j] = *avg_Kl_div*(SEMANTIC_RELATION[i], SEMANTIC_RELATION[j])
Return DISTANCE_MAT

IV. Explainable Semantic Relations in Sentinel-1 Images Using LDA

A. Experimental Procedure

The supervised classification of *sea-ice* images from [15] (used as a case study) shows the semantic classes of the dataset. We used these classes as a reference for finding semantic relations using intermediate representations (topics) using an LDA model. Since the last decade, LDA has been successfully used in remote sensing image understanding. LDA originated as a generative probabilistic model of text data and works in a bag-of-words framework. However, our approach focuses on using LDA as a seamless and explainable data mining tool to discover the semantic relations in Sentinel-1 images.

The bag-of-words framework is a prerequisite for applying LDA. In general, the steps in the bag of words modeling in the image processing domain are patch sampling, local feature extraction, dictionary learning, word assignment, and histogram computation. There are two parameters to be determined for patch sampling: the patch size and the sampling strategy. It has been shown that a smaller patch size yields better classification accuracies both for overlapping and nonoverlapping patches [18]. Furthermore, one has to decide upon a sampling technique to sample p patches from the image dataset, for instance, random sampling or dense sampling. It was shown that random sampling and dense sampling do not cause a big difference in the classification accuracy [18]. The next step is to extract x local descriptors from the p sampled collection of patches. In case of no features, the feature vectors are the vectorized pixel values of the patches. The third step is learning a dictionary $D = (d_1, \ldots, d_v)$ with V words using all local features from the p patches. Usually, this is done by an unsupervised learning method, e.g., k means clustering or a Gaussian mixture model. Each entry d_i in the dictionary is the center of a cluster. The next step is to find a dictionary-based representation $r = [r_1, \ldots, r_v]$ for each previously extracted local descriptor x and its dictionary size of v words. This can be done using a hard feature assignment or a soft assignment [18] to each local descriptor x. Finally, r, the descriptor representation has only one nonzero element. Next, for each image, the count of each word is computed to get the final bag of words model of the image dataset.

LDA is a probabilistic model for discrete data, which discovers latent semantics of the documents as a set of topics, where the topics are distributions over the words of the dictionary. To apply LDA to images, we need to find the analogies to *corpora*, *documents*, topics, *dictionaries*, and *words*.

In our approach of applying LDA to Sentinel-1 images, each scene I_i ($i = 1, \ldots, N$) is considered as a corpus. M patches (henceforth referred to as macropatches) of size (H_i, W_i) pixels are extracted from each scene to serve as visual documents using the dense sampling.

We chose a fixed size for these documents, which is the same as the patch size of the reference research. The words are obtained by the quantization of local image features. The vectorized pixel values of micropatches (extracted from the macropatches) are then used as local features. Each document (macropatch) M_i consists of N_i visual words (micropatch). A dictionary of size V is computed by the vector quantization of the feature space using a k means clustering approach. For LDA, the number of topics K must be specified. We set the number of topics K to roughly match the number of classes in the reference research. This concept is presented in Fig. 4.

1) LDA Generative Process: The following generative process (see Fig. 4) is pursued to model M documents, each document containing N words and K topic*s*:

For each document d_i*:*

1) Choose $\theta_i \sim$ Dirichlet (α), $i \in \{1, \ldots, M\}$;
2) Choose $\varphi_k \sim$ Dirichlet (β), $k \in \{1, \ldots, K\}$.

For each word position w_{ij} in each image patch i:

1) Choose a topic $z_k \sim$ multinomial (θ_i);
2) Choose a word $w_{i,j} \sim$ multinomial ($\varphi_{z_{i,j}}$)

where α and β are the Dirichlet parameters determining the sparsity of the document-topic and of the topic-word distributions. Given the α and β parameters and the number of topics K, LDA models the documents as the probability distributions of the topics.

The joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_k \mid \theta) p(w_n \mid z_n, \beta) \quad (1)$$

where $p(z_n|\theta)$ is actually the θ_i for the unique i such that $z_n^i = 1$. Integrating over θ and summing over z, we obtain the marginal distribution of a document

$$p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) \times \left(\prod_{n=1}^{N} \sum_{z_n} p(z_{dn} \mid \theta)\ p(w_n z_n, \beta) \right) d\theta. \quad (2)$$

We retrieve the document-topic probability matrix $\theta_{M \times K}$ from the trained model. Each vector θ_i ($i = 1, 2, \ldots, M$) is a K-size vector containing the topic probability distribution for document d_i, i.e.,

$$\theta_{ik} = p(z_k = 1 \mid d_i = 1). \quad (3)$$

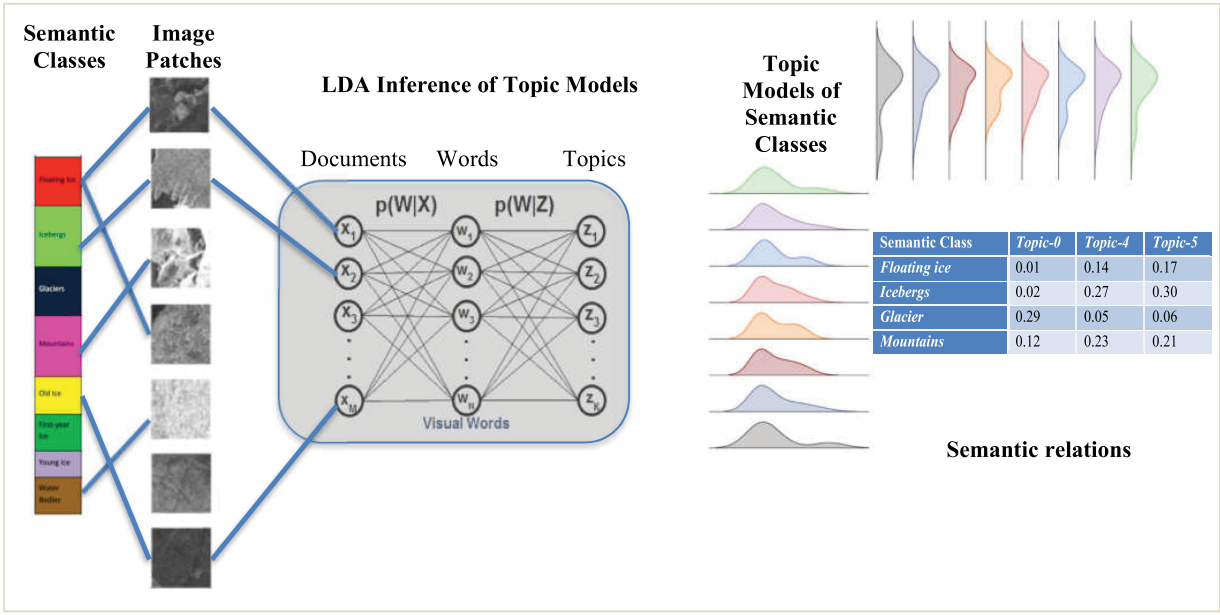
Likewise, LDA models the topics as distributions over word probabilities and outputs another matrix φ where

$$\Phi_{kj} = p(w_j = 1 \mid z_k = 1). \quad (4)$$

2) Semantic Relations: After modeling a visual corpus with LDA, we get the topic-word probability matrix $\theta_{M \times K}$. This matrix is then used to assign the maximum probable visual topic to each visual word in each visual document according to

$$\text{Word } j \rightarrow \max(p(z_k \mid w_j). \quad (5)$$

This, in turn, labels the micropatches used for building visual words with visual topics. Such a representation now allows one to compute the probability of each visual topic for each reference semantic class as the count of micropatches being labeled with a visual topic, divided by the total number of micropatches labeled with any visual topic, i.e., the probability of a visual topic t_i for



Semantic Class	Topic-0	Topic-4	Topic-5
Floating ice	0.01	0.14	0.17
Icebergs	0.02	0.27	0.30
Glacier	0.29	0.05	0.06
Mountains	0.12	0.23	0.21

Fig. 4. Topics as an intermediate level of the information modeled by LDA.

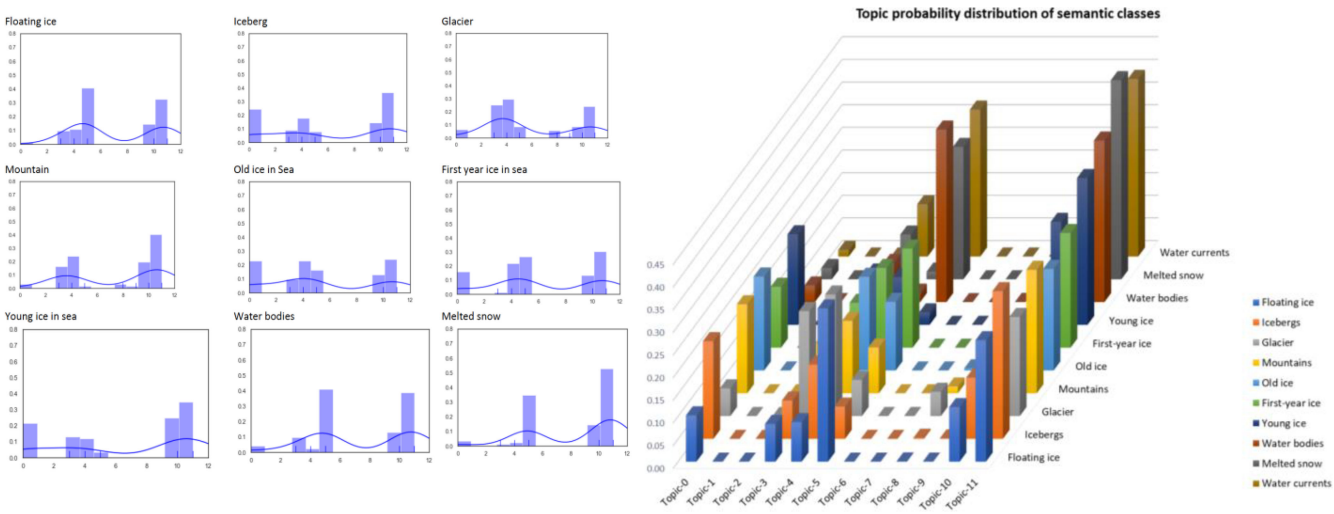


Fig. 5. Semantic relations derived by using topic distributions.

a reference semantic class C_j is computed as

$$p\left(t_i \mid C_j\right) = \text{Count}\left(\text{Micropatch}\left(t_i\right)\right) \mid \times \sum_{0}^{K-1} \text{Count}\left(\text{Micropatch}\left(t_i\right)\right). \quad (6)$$

For each reference class C_i ($i = 1, 2, \ldots,$ Num_of_classes), the semantic relation is heuristically stated as

$$C_i = \sum_{k} p\left(t_k\right) * t_k + ui \quad (7)$$

where i is the index for each semantic class, k is the index for a topic, $p(t_k)$ denotes the probability of the kth visual topic in class C_i, while u_i is the normalized count of words with no assigned topic to class C_i. These semantic relations are scientifically explainable by domain experts.

B. Experimental Results

We followed the experimental procedure as described in Section II to produce the results. Then the semantic relations of the 11 classes (including black edges) were computed and shown in Fig. 5. We did not consider a black edge (image edge effects) as a distinct class and present only the semantic relations of the other ten classes.

C. Explainable Outcomes

The topics retrieved by LDA are considered to be in the interpretable domain, because, when visually presented and

compared with the classification maps or combined with the domain knowledge, they are able to generate some interpretation. The features in the interpretable domain are used to provide explanations. However, semantic relations computed by using the probability of topics are only explainable by domain experts.

The semantic relations of each reference class indicate the physical composition of that particular class, which means that a domain expert can specify the nature of each topic. Although the visual topics are modeled as the distributions of visual words (which are vectorized pixel values and have no semantic meaning), an explanation by a domain expert would assign semantic meanings to the visual topics. This domain expert can also give scientific reasons for the presence, absence, and abundance of particular visual topics in each particular class. This offers a way of how to generate intermediate semantic labels from low-level features, such as pixel values, and to explain them.

The semantic classes floating ice, water bodies, melted snow/ice, and water/ice currents contain mostly topics Topic-5 (0.34, 0.38, 0.29, and 0.32) and Topic-11 (0.27, 0.36, 0.44, and 0.39). We also found that the topic distributions of each class vary every month due to the varying characteristics of the land-cover classes (not presented here). This allows us not only to identify the observation month but also to conduct time-series analyses of land-cover classes. In Fig. 5, we present the semantic relations for the given classes in terms of topic probabilities. The floating ice class is dominated by Topic-5 and Topic-11, whereas the icebergs class has less appearance of Topic-5 but more of Topic-0. Topic-11 is dominant in all classes but this is not the case for Topic-0, Topic-3, Topic-4, Topic-5, and Topic-10. Here, a domain expert can explain the nature of the topics, e.g., Topic-5 and Topic-11, and even nonexperts can intuitively confirm that Topic-5 is a water-like element.

V. Interpretable Topic Representations Using LDA

A. Experimental Procedure

Topic representation maps can be created easily by using the topic-word probability matrix retrieved from LDA. Each visual word in the visual corpus is assigned to the most probable topic according to

$$\text{Word } j \rightarrow \max(p(z_k \mid w_j). \tag{8}$$

The topic representation map can be visualized by drawing each micropatch filled with a color identifying the labeling topic.

Thus, drawing a topic representation map for each visual corpus recreates each scene with colors identifying the visual topics. This map gives a visual impression of the nonvisual Sentinel-1 information and shows the abundance of the visual topics, which enables us to make guesses about the study area with limited domain knowledge.

B. Experimental Results

Each visual word is a 4×4 pixel cut out from the scene and by coloring each cut out with the specific color that identifies its most probable topic, we obtain a topic representation map that recreates the scene in a completely unsupervised manner. According to the annotations, 11 classes (including the black edges) could be found in the images comprising floating ice, glaciers, icebergs, melted snow/ice, mountains, old ice, first-year ice, young ice, water bodies, and water/ice currents.

The left column of Fig. 7 shows the original quick-look data of an image acquired in August 2018, while the central column represents the semantic annotations obtained from an active learning framework [15], and the right column presents the topic representations. Fig. 7 also displays the quick-look images and the topic representations for two images taken in June and April 2018.

C. Seamlessness to Interpretability

In our research, the topics retrieved from the LDA model are the features in the interpretable domain. We used these topics to create visualizations called "topic representation map" for subsequent interpretation. Thus, our method proposes a way to map seamless features into interpretable maps. Unlike a color slide, their interpretation requires some domain knowledge. In the following (see Fig. 6), we present the three case studies from our research as follows.

1) Case study-1 demonstrates no or very limited interpretability, using only the topic representation map.
2) Case study-2 shows their limited interpretability, achieved by comparing the topic representation map and the corresponding product quick-look image.
3) Case study-3 is done by comparing the topic representation map with the corresponding classification map; this produces a great amount of interpretable information.

We observe that a greater amount of interpretability is achieved with a higher quality of the domain knowledge. However, one can see that the topic representation map gives more granular details than the quick-look image (case study-2) or the reference classification map (case study-3).

We define three levels of interpretability with respect to our study, as depicted in Fig. 6(d). Low interpretability means only being able to find the shapes of objects, the homogeneity of topic compositions together with the spatial relations of these objects, and the topics composing these objects. Medium interpretability is defined as all interpretabilities defined in low interpretability plus a matching of brightness levels. Finally, high interpretability refers to medium interpretability plus a visual class-topic mapping. All these levels of interpretability can be demonstrated in three case studies.

1) Case Study-1: Low Interpretability: From the topic representation map of the study area, acquired on August 9, 2018, we can only see the shapes of objects composed of a set of topics, the spatial relations of the objects, and the homogeneity of the topics composing the objects. Some objects are made of homogeneous topics, while others are heterogeneous combinations of topics. For instance, the area, as shown in Fig. 6(a), is mostly described by the cyan topic, with a significant amount of dark green and pink topics. We also find some dark green chunks, having some pink objects on top, surrounded by almost homogeneous cyan-colored objects. This method offers a limited amount of interpretability, denoted as low interpretability.

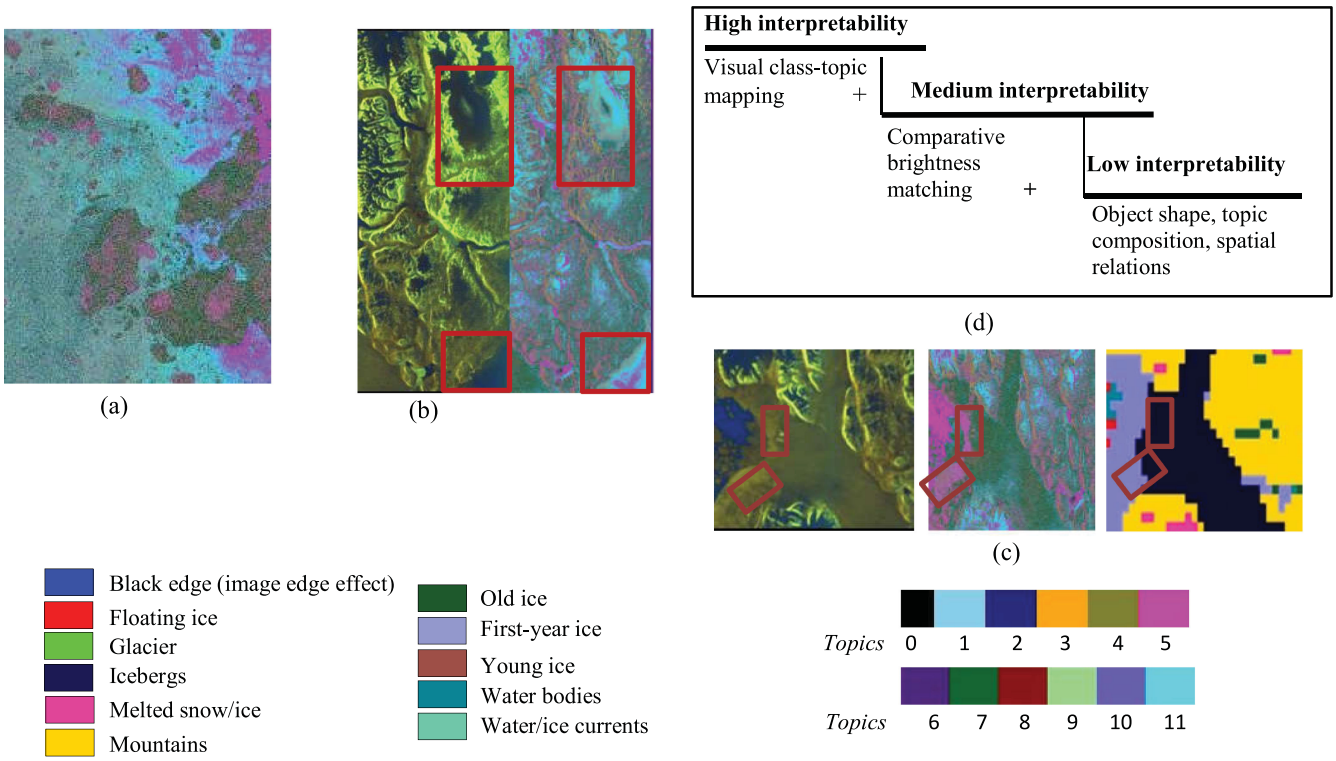


Fig. 6. Interpretation of topic representation maps. (a) Case study-1. (b) Case study-2. (c) Case study-3. (d) Levels of interpretation in our study. The bottom right color legend is linked to the semantics retrieved by the active learning method, while the bottom left color legend is linked to the topic distribution.

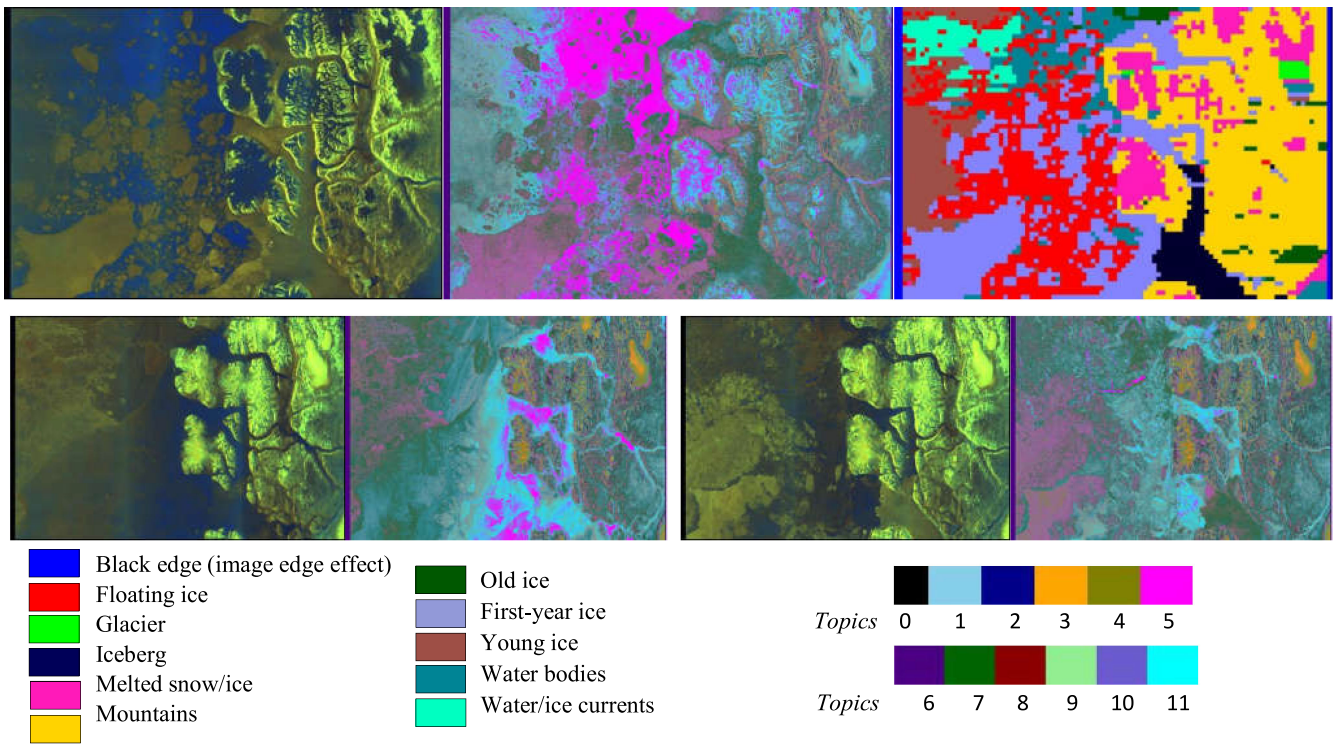


Fig. 7. Topic representation maps. (Top from left to right) A quick-look of Sentinel-1 image acquired in April 2018, followed by its LDA topic representation, and the semantic map of the active learning step. (Middle from left to right) The first two left images are the quick-look of Sentinel-1 image from June 2018 together with its LDA topic representations, while the last two right images are the quick-look of Sentinel-1 image from August 2018 and its LDA topic representation map. (Bottom from left to right) The left color legend is the one for the semantic classes of the active learning, while the right color legend illustrates the topic distribution.

2) Case Study-2: Medium Interpretability: A visual comparison of the topic representation maps with the product quick-look images can provide a good guess of the study area. Here, we use the quick-look image as a source of a limited amount of domain knowledge, which provides information about the spatial relations of the objects in the scene and their comparative brightness. We can easily find the mapping between objects in the quick-look image and the objects in the topic representation map. The red boxes in Fig. 6(b) show the shape of the objects, the homogeneity of the topics, and their mutual spatial relations. The darkest objects in this quick-look image appear as a cyan topic in the topic representation map. The objects with a higher brightness have the dark green topic assigned to them. The objects in the lowest right area map to the objects with a cyan topic, with some appearance of pink topics in the topic representation map, which may stand for another level of brightness, not being visible in the quick-look image. Following this approach, it is possible to make consistent interpretations over the whole scene, also with greater granular details.

3) Case Study-3: High Interpretability: By using a quick-look image and its reference classification map as the sources of domain knowledge, we can achieve a greater level of interpretability, which gives us information about the shape of the objects, the homogeneity of topic compositions, the mutual spatial relations of objects and topics, the brightness differences, and our heuristic class-topic mapping.

A class-topic mapping can be interpreted by a visual comparison of a topic representation map with its reference classification map. For instance, Fig. 6(c) shows an iceberg area from the quick-look image, its topic representation, and the resulting classification map. From this, we can interpret the shape of the object "icebergs" object, that icebergs are mostly linked with the green topic, with some pink patches of this topic appearing in the portions that look brighter in the quick-look image, and that the pink topic patches appear around the left boundary of the iceberg object. Noticeably, we obtain an additional granularity of information that is missing in the classification map but is visible in the topic representation map.

Furthermore, Figs. 6 and 7 also show another example of interpretation, this time without the classification map.

VI. Explainable Class Distinctions

A. Experimental Procedure

Once the semantic relations for each class have been computed, we get a signature for each class in terms of the probabilities of the visual topics. Each term in the signature is the probability of a topic, followed by the id of that topic and an $*$ operator. The general formula of the semantic classes can be heuristically written as

$$C_i = \sum_k p\ (t_k) * t_k + ui \tag{9}$$

where i is the index for each semantic class, k is the index for a topic, $p(t_k)$ denotes the probability of the kth topic in class C_i, and u_i is the normalized count of words with no assigned topic in class C_i, while u_i is the normalized count of words with no assigned topic in class C_i.

This gives a probability distribution for each class. Now it is possible to distinguish between the classes to see whether that gives some scientifically consistent decision. This step also serves as a validation of the approach. To distinguish between the classes, we have to distinguish between the probability distributions. The Kullback–Leibler (K–L) divergence is the state-of-the-art method for this. The divergence from a distribution Q to a distribution P is defined as

$$D\mathrm{KL}\left(P \mid\mid Q\right) = \sum \left(P\left(x\right)\right)\ \log\left(\frac{P\left(x\right)}{Q\left(x\right)}\right). \tag{10}$$

However, for using the K–L divergence as a distance metric, we need to have symmetric values

$$\mathrm{dist}_{i,j} = \mathrm{dist}_{j,i\ ,i\neq j.} \tag{11}$$

To have a symmetric form of the K–L divergence, we averaged the divergence between every two classes and used these values as a distance metric, which results in a symmetric matrix of interclass distances. Next, to have a clear grouping of similar and dissimilar classes, in terms of this distance metric, we subjected this distance matrix to a hierarchical clustering algorithm. Hierarchical clustering is a common analysis tool for relationship discovery based on some distance metric.

B. Experimental Results

The results of this step are obtained by subjecting the 11×12 semantic relation matrix (11 classes and 12 topics) to the K–L divergence computation module to retrieve the distance matrix. We get an 11×11 symmetric matrix. This distance matrix is shown in Fig. 8. The distances of less than the mean value of all distances are highlighted in light green, the remaining ones are highlighted in cyan, while the diagonal elements are set to zero and highlighted in dark green. We get an initial grouping of *similar* and *dissimilar* classes from this. For a better classification, based on the measure of similarity using the K–L divergence, we ran a hierarchical clustering on the classes, thus, grouping them into clusters. Although it needs a domain expert to confirm the correctness of the grouping, we can claim that the resulting dendrogram (see Fig. 9) reflects our intuition about the physical properties of the semantic classes. The set of classes floating ice, water bodies, melted snow, and water currents form the first group in the dendrogram (see Fig. 9). This is distinguished from the second group containing glacier, young ice, icebergs, mountains, old ice, and first-year ice.

Looking closer into the first group, we see that floating ice is separated from the subgroup containing water bodies, melted snow, and water currents.

The grouping of land-cover classes and their topic compositions are further summarized in Fig. 9. This actually shows a cause–effect relationship between class groupings and topic compositions. In other words, this tells us which classes are closer to each other and why, as well as which classes are distant from each other, and the reason behind them for being so distant.

	Floating ice	Icebergs	Glacier	Mountains	Old ice	First-year ice	Young ice	Water bodies	Melted ice	Water currents
Floating ice	0	0.28	0.53	0.25	0.17	0.07	0.47	0.1	0.16	0.23
Icebergs	0.28	0	0.32	0.05	0.06	0.12	0.07	0.57	0.55	0.74
Glacier	0.53	0.32	0	0.2	0.27	0.25	0.5	0.81	0.73	1.04
Mountains	0.25	0.05	0.2	0	0.05	0.08	0.15	0.52	0.5	0.72
Old ice	0.17	0.06	0.27	0.05	0	0.03	0.21	0.5	0.55	0.74
First-year ice	0.07	0.12	0.25	0.08	0.03	0	0.29	0.3	0.34	0.48
Young ice	0.47	0.07	0.5	0.15	0.21	0.29	0	0.71	0.63	0.77
Water bodies	0.1	0.57	0.81	0.52	0.5	0.3	0.71	0	0.03	0.03
Melted ice	0.16	0.55	0.73	0.5	0.55	0.34	0.63	0.03	0	0.03
Water currents	0.23	0.74	1.04	0.72	0.74	0.48	0.77	0.03	0.03	0

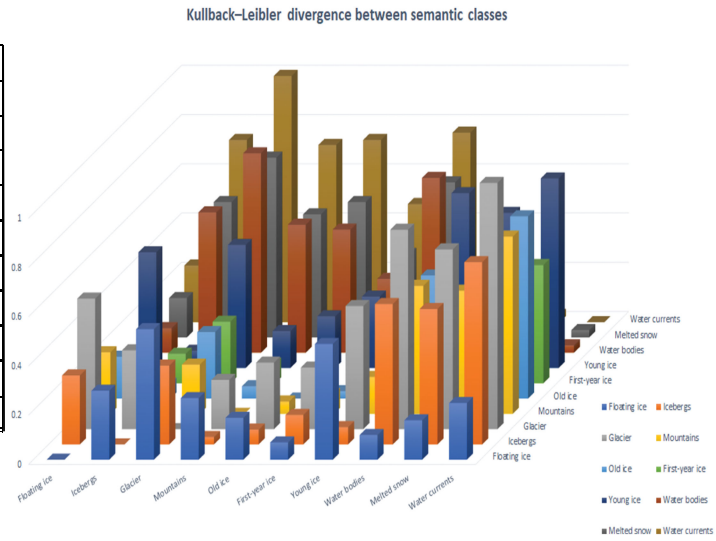


Fig. 8. Class distinction using the K–L divergences.

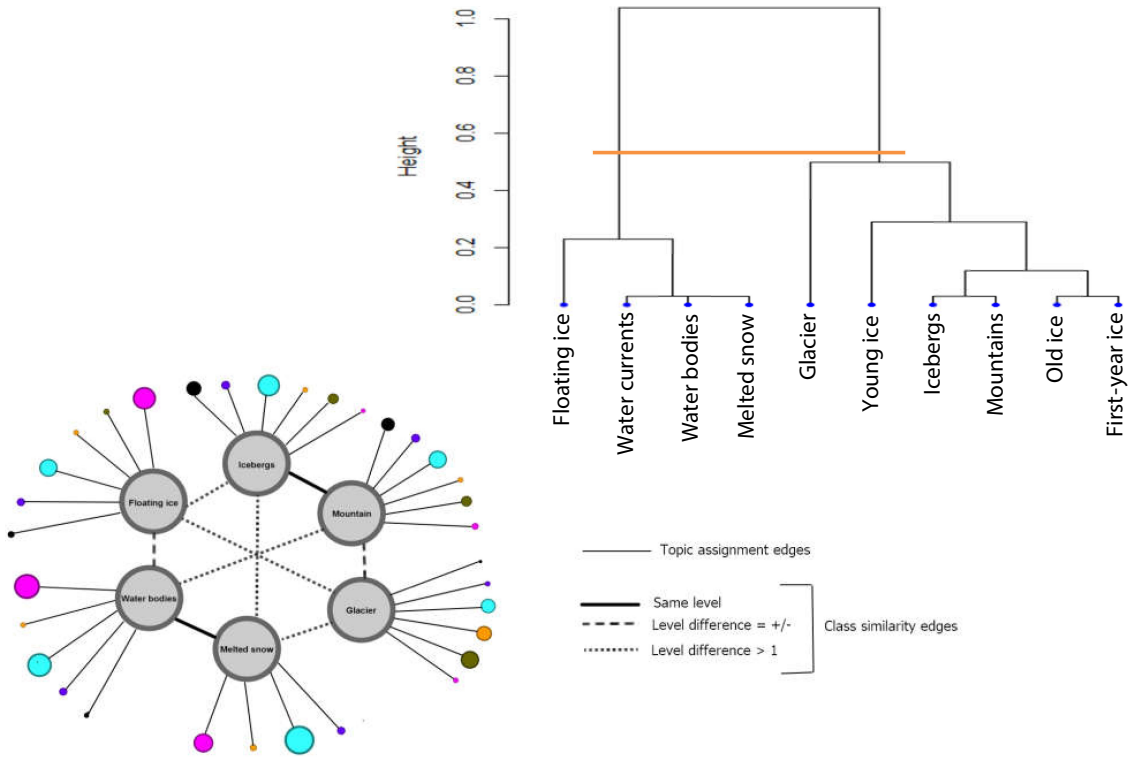


Fig. 9. Grouping of similar classes based on their semantic relations.

The graph shown in Fig. 9 contains two types of nodes and two types of edges as follows.

1) *Labeled nodes:* These nodes depict the land/surface cover class. For visual simplification, only six classes are chosen for display.
2) *Unlabeled colored nodes*: For each class, they indicate the topic probabilities for six topics. The larger the node, the higher is the probability of that particular topic.
3) *Topic assignment edges*: They relate each class to its component topics identified by colored nodes.
4) *Class similarity edges*: These edges are of three types. Each type identifies the group relationship between the pair of class nodes they join. These relationships are the level differences obtained when the dendrogram is considered as a tree structure.

We see that the nodes, water body and melted snow (depicting the classes water bodies and melted snow), both have large cyan and pink nodes indicating that both of them are dominated by cyan and pink topics, and consequently are on the same level of the dendrogram. The bar graph presents the dissimilarity of

the semantic classes. We can track each class and describe its dissimilarity from the other classes by following each row in the bar graph. For instance, the floating ice class is highly dissimilar with glacier and young ice but very similar to water bodies. Repeated experiments found similar groupings of classes, thereby ensuring our scientific and algorithmic consistency.

C. Explainable Outcomes

The semantic relations between reference classes and visual topics define the class compositions in terms of topics. From the distance matrix shown in Fig. 8, we can get an initial idea about the similarity and the dissimilarity between each pair of semantic classes. For instance, the glacier class is similar to mountains, old ice, first-year ice, and young ice. We can algorithmically explain these phenomena in terms of the K–L divergence.

The K–L divergence compares the abundance of each visual topic between two given classes and finds the smallest divergence between glacier (a $*$ Topic-0 $+$ b $*$ Topic-1 $+$ c $*$ Topic-3 $+\ldots$) and mountains (a $*$ Topic-0 $+$ b $*$ Topic-1 $+$ c $*$ Topic-3 $+\ldots$), the second smallest divergence between glacier and young ice (a $*$ Topic-0 $+$ b $*$ Topic-1 $+$ c $*$ Topic-3 $+\ldots$), whereas the glaciers class has the greatest divergence between glacier and melted snow/ice (a $*$ Topic-0 $+$ b $*$ Topic-1 $+$ c $*$ Topic-3 $+\ldots$) because they have more dissimilar topic probability distributions.

However, these interclass distances are scientifically explainable by domain experts who can comment on the physical similarity of the classes and map that information to our findings.

Consequently, the results of the hierarchical clustering based on the similarity of such compositions provide explainable outcomes.

We found that the classes that are physically similar, for example, water body, melted snow, and water currents are dominated by the same set of topics. The water body class is composed of Topic-5 and Topic-11. The same holds for the classes melted snow and water currents, the only difference being the amounts of Topic-3 and Topic-4, which are higher in the latter two cases.

A dendrogram is actually a tree representation of the results of hierarchical clustering. Cutting a given tree at various height levels along the *y*-axis helps understand the similarity of the classes. If we cut our tree at 0.5, we are left with two clusters. Starting from the left, we first get floating ice, water currents, water bodies, and melted snow. Then the second cluster contains the classes glacier, young ice, icebergs, mountains, old ice, and first-year ice.

When we look at the first cluster obtained by the cut at 0.5, this cluster contains floating ice, water currents, water bodies, and melted snow. In this cluster, the floating ice class stands apart from the other three classes (water body, melted snow, and water currents), which are most similar to each other. Looking closer into the second cluster obtained by the cut at 0.5, we see that the glacier class is an outlier. When we apply a cut at a height of 0.3, we are left with a lone cluster glacier, and another one containing young ice, icebergs, mountains, and first-year ice. Among these four, obviously, the class first-year ice is the most distant one. Among the other four, iceberg and mountains are most similar to each other and different from the other group of similar classes, namely, old ice and first-year ice.

Any dissimilarity of classes can also be understood from the diagram. From a physical perspective, the classes, mountains and water bodies, must be distinct. Such a decision can also be reached from the topic distributions of these two classes. The mountains class has a fair share of Topic-0, Topic-3, Topic-4, Topic-5, and a large amount of Topic-11.

In the case of the class water body, there is much less appearance of other topics, except for Topic-5 and Topic-11.

VII. Conclusion and Future Work

The idea behind this work is to apply unsupervised topic models, such as LDA, as a data mining tool for learning high-level semantic structures in image areas with no or poor existing prior knowledge. We demonstrated our approach using Sentinel-1 data for an area near the North Pole for which no ground-truth data existed.

LDA, when being used as a data mining tool, maps each semantic class into combinations of topics and demonstrates the method's capability to detect new semantics. Here, the images are represented as the distributions of the topics. We further used a sample annotation of the images, derived from and based on the results of an active learning method to investigate the relationships between the topics and the land/surface cover classes. To this end, the topic distribution of each class was computed. The results show that different classes have different topic distribution signatures.

In addition, to use all information provided by Sentinel-1 data, we also considered the combination of its two polarizations [17], and we compared the results of the average (avg) and the difference of polarizations with those obtained by the HH polarization. Three sets of experiments are performed based on the combinations of polarization, i.e., HH, HH-HV, combination of HH, HV, and avg (HH+HV). The obtained results show a consistent structure and also similar class groupings using the K–L divergence on semantic relations. However, more visual granularity is obtained by combining polarizations [HH, HV, avg (HH, HV)].

Furthermore, LDA is a well-studied subject; however, the uniqueness of our approach lies in the following aspects.

A. Featurelessness

As there is no prior knowledge about the dataset, we use a seamless data mining method to avoid any bias caused by the feature model. The pixel brightness values of the image patches are used as features for the machine learning model.

B. Unsupervised Discovery of the Innate Data Model Ranked by Trust

The model is ranked by trust levels of the topic probabilities. For each word in the bag of words framework, the maximum probable topic given by the LDA model is chosen; this, in turn, removes the irrelevant topics and leads to a refined result.

C. Visual Interpretability

The topic representation map provides a way to visually interpret the study area by comparing the map with the product quick-look image, without requiring any reference data.

D. Explainability and a New Validation Method

We used the topics as data model descriptors to provide an explanation of the composition of the land/surface cover classes. We quantitatively determined their semantic relations, using the topics as an intermediate-level data representation model.

In addition, we proposed methods dedicated to exploring unknown data without training or verification information. The validation of results obtained when using LDA is usually made with the classification results and using some reference dataset. In our case, we applied a unique validation method that computes the interclass dissimilarities by exploiting the K–L divergence between the topic probability distributions of the land/surface cover classes.

E. Visual Representation From the Nonvisual Sentinel-1 Images

One of the outcomes of our research is a visual representation of selected areas covered by Sentinel-1 images. When scientifically explained, they can serve as a benchmark dataset for further research and/or applications. This idea is planned as future work.

The data of this study can be requested from the authors under the conditions imposed by the projects that funded this study.

ACKNOWLEDGMENT

The authors would like to thank R. Bahmanyar for his support.

REFERENCES

[1] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation'," *AI Mag.*, vol. 38, no. 3, pp. 50–57, 2017.
[2] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019.
[3] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, 2018.
[4] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[5] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42200–42216, 2020.
[6] R. Fernandez-Beltran, J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "Remote sensing image fusion using hierarchical multimodal probabilistic latent semantic analysis," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4982–4993, Dec. 2018.
[7] M. Lienou, H. Maitre, and M. Datcu, "Semantic annotation of satellite images using latent Dirichlet allocation," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 28–32, Jan. 2010.
[8] D. Espinoza-Molina, R. Bahmanyar, M. Datcu, R. Díaz-Delgado, and J. Bustamante, "Land-cover evolution class analysis in image time series of Landsat and Sentinel-2 based on latent Dirichlet allocation," in *Proc. 9th Int. Workshop Anal. Multitemporal Remote Sens. Images*, Brugge, Belgium, 2017, pp. 1–4.
[9] C. Vaduva, I. Gavat, and M. Datcu, "Latent Dirichlet allocation for spatial analysis of satellite images," *IEEE Trans.Geosci. Remote Sens. Lett.*, vol. 51, no. 5, pp. 2770–2786, May 2013.
[10] R. Tănase, R. Bahmanyar, G. Schwarz, and M. Datcu, "Discovery of semantic relationships in PolSAR images using latent Dirichlet allocation," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 237–241, Feb. 2017.
[11] D. Bratasanu, I. Nedelcu, and M. Datcu, "Bridging the semantic gap for satellite image annotation and automatic mapping applications," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 4, no. 1, pp. 193–204, Mar. 2011.
[12] Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, vol. 61, no. 10, pp. 36–43, 2018.
[13] A. Förster, J. Behley, J. Behmann, and R. Roscher, "Hyperspectral plant disease forecasting using generative adversarial networks," in *Proc. Int. Geosci. Remote Sens. Symp.*, 2019, pp. 1793–1796.
[14] S. Ghosal, D. Blystone, A. K. Singh, B. Ganapathysubramanian, A. Singh, and S. Sarkar, "An explainable deep machine vision framework for plant stress phenotyping," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 18, pp. 4613–4618, 2018.
[15] C. O. Dumitru, V. Andrei, G. Schwarz, and M. Datcu, "Machine learning for sea ice monitoring from satellites," in *Proc. Munich Remote Sens. Symp.*, 2019, pp. 83–89.
[16] J. Haarpaintner, R. T. Tonboe, D. G. Long, and M. L. V. Woert, "Automatic detection and validity of the sea-ice edge: An application of enhanced-resolution QuikScat/SeaWinds data," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1433–1443, Jul. 2004.
[17] S. Abdikan, F. B. Sanli, M. Ustuner, and F. Calò, "Land cover mapping using Sentinel-1 SAR data," in *Proc. 23rd ISPRS Congr.*, Prague, Czech Republic, 2016, pp. 757–761.
[18] S. Cui, G. Schwarz, and M. Datcu, "Remote sensing image classification: No features, no clustering," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 8, no. 11, pp. 5158–5170, Nov. 2015.
[19] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim, "Towards a rigorous evaluation of XAI methods on time series," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, Seoul, Korea, 2019, pp. 4197–4201.
[20] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 8, pp. 2674–2693, Aug. 2019.
[21] Z. Xi and G. Panoutsos, "Interpretable machine learning: Convolutional neural networks with RBF fuzzy logic classification rules," in *Proc. Int. Conf. Intell. Syst.*, Funchal, Portugal, 2018, pp. 448–454.
[22] F. Wick, U. Kerzel, and M. Feindt, "Cyclic boosting—An explainable supervised machine learning algorithm," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl.*, Boca Raton, FL, USA, 2019, pp. 358–363.
[23] C. O. Dumitru and M. Datcu, "Information content of very high-resolution SAR images: Study of feature extraction and imaging parameters," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 8, pp. 4591–4610, Aug. 2013.
[24] C. O. Dumitru, G. Schwarz, and M. Datcu, "SAR image land cover datasets for classification benchmarking of temporal changes," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1571–1592, May 2018.
[25] L. Jiao, X. Tang, B. Hou, and S. Wang, "SAR images retrieval based on semantic classification and region-based similarity measure for earth observation," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 3876–3891, Aug. 2015.

Chandrabali Karmakar received the M.Sc. degree in forest information technology from Eberswalde University for Sustainable Development, Eberswalde, Germany, in 2017.

Since 2019, she has been a German Academic Exchange Service (DAAD) Ph.D. Fellow with German Aerospace Center, Cologne, Germany, under the supervision of Dr. M. Datcu. She has several years of work experience in the information technology industry, as a Software Developer. Her research interests include explainable artificial intelligence, data mining, and remote sensing.

Corneliu Octavian Dumitru received the B.S. and M.S. degrees in applied electronics from the Faculty of Electronics, Telecommunications and Information Technology, and the Ph.D. degree in engineering, all from Politehnica University of Bucharest, Bucharest, Romania, in 2001, 2002, and 2006, respectively, and the Ph.D. degree in telecommunications from Sorbonne University Pierre and Marie Curie Campus, Paris, France, in 2010.

From 2005 to 2006, and in 2008, he was a Coordinator for two national grants delivered by the Romanian Ministry of Education and Research. Since 2010, he has been a Scientist with the Remote Sensing Technology Institute, German Aerospace Center, Oberpfaffenhofen, Germany. With Politehnica University, he had a teaching activity as a Lecturer, delivering lectures and seminaries and supervising laboratory works in the fields of information and estimation theory, communication theory, and signal processing. Since 2004, he has been cosupervising bachelor, master, and Ph.D. theses. He is currently involved in several projects in the frame of the European Space Agency and European Commission Programmes for information extraction, taxonomies, and artificial intelligence/machine learning and knowledge discovery using remote sensing imagery. His research interests include stochastic process information, model-based sequence recognition and understanding, basics of man–machine communication, information management, semantics, data mining, and image retrieval in extended databases.

Gottfried Schwarz received the graduate degree in engineering from the Technical University of Munich, Munich, Germany, in 1976.

Since many years, he has been involved in a number of national and international space projects with the German Aerospace Center, Oberpfaffenhofen, Germany; among them were deep-space missions as well as Earth observation missions. In particular, he has been involved in the design of deep-space instruments from initial engineering studies to detailed design work, modeling of instrument performance, instrument assembly and testing, real-time experiment control, instrument checkout and calibration, data verification and validation, as well as data processing and scientific data analysis. Besides instrument-related aspects, he has also many years of experience in the processing and analysis of various instrument data within ground segments, in particular of optical and SAR remote sensing data, in the interpretation of geophysical data with emphasis on retrieval algorithms with forward modeling, inversion techniques, and data mining. He also has special experience in signal processing resulted from engagement in image data compression and feature analysis together with the performance analysis of image classification.

Mihai Datcu (Fellow, IEEE) received the M.S. and Ph.D. degrees in electronics and telecommunications from the University Politechnica of Bucharest (UPB), Bucharest, Romania, in 1978 and 1986.

In 1999, he received the title Habilitation à Diriger des Recherches in computer science from Louis Pasteur University, Strasbourg, France. He is currently a Senior Scientist and Image Mining Research Group Leader with the Remote Sensing Technology Institute, German Aerospace Center, Cologne, Germany, and a Professor with the Department of Applied Electronics and Information Engineering, Faculty of Electronics, Telecommunications, and Information Technology, UPB, Bucharest. From 1992 to 2002, he had a longer Invited Professor assignment with the Swiss Federal Institute of Technology, ETH Zurich. From 2005 to 2013, he was the Professor Holder of the DLR-CNES Chair with ParisTech, Paris Institute of Technology, Telecom Paris. His interests are in data science, machine learning and artificial intelligence, and computational imaging for space applications. He is involved in big data from Space European, ESA, NASA, and national research programs and projects.

Dr. Datcu was a recipient of the Best Paper Award of the IEEE Geoscience and Remote Sensing Society, in 2006. He is a member of the ESA Big Data from Space Working Group and the holder of a 2017 Blaise Pascal Chair at CEDRIC, CNAM, France.