



Interpretable land cover classification with modal decision trees

G. Pagliarini & G. Sciavicco

To cite this article: G. Pagliarini & G. Sciavicco (2023) Interpretable land cover classification with modal decision trees, European Journal of Remote Sensing, 56:1, 2262738, DOI: [10.1080/22797254.2023.2262738](https://doi.org/10.1080/22797254.2023.2262738)

To link to this article: <https://doi.org/10.1080/22797254.2023.2262738>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 18 Dec 2023.



Submit your article to this journal [↗](#)





View related articles [↗](#)



View Crossmark data [↗](#)

Interpretable land cover classification with modal decision trees

G. Pagliarini ^{a,b} and G. Sciavicco ^a

^aDepartment of Mathematics and Computer Science, University of Ferrara, Ferrara, Italy; ^bDepartment of Mathematical, Physical and Computer Sciences, University of Parma, Parma, Italy

ABSTRACT

Land cover classification (LCC) refers to the task of classifying each pixel in satellite/aerial imagery by predicting a label carrying information about its nature. Despite the importance of having transparent, symbolic decision models, in the recent literature, LCC has been mainly approached with black-box functional models, that are able to leverage the spatial dimensions within the data. In this article, we argue that standard symbolic decision models can be extended to perform a form of spatial reasoning that is adequate for LCC. We propose a generalization of a classical decision tree learning model, based on replacing propositional logic with a modal spatial logic, and provide a CART-like learning algorithm for it. We evaluate its performance at five different LCC tasks, showing that this technique leads to classification models whose performances are superior to those of their propositional counterpart, and at least comparable with those of non-symbolic ones. Ultimately, we show that spatial decision trees and random forests are able to extract complex, but interpretable spatial patterns.

ARTICLE HISTORY

Received 10 December 2022
Revised 4 September 2023
Accepted 19 September 2023

KEYWORDS

Interpretable machine learning; modal logic; decision tree learning; hyperspectral image classification; land use; land cover

Introduction

Connectionist learning has been the driving force of machine learning for at least a decade, allowing great advances in many difficult artificial intelligence (AI) tasks. Recent trends, however, have seen a rising need for models that are transparent and interpretable, and it has been pointed out that, despite their potential and versatility, neural models often enclose complex mathematical functions that are difficult to interpret; such a need is even reflected in recent political initiatives (see, e.g. the 2016 General Data Protection Regulation of the European Union). A well-known approach to interpretable modeling dates back to the beginning of AI and is referred to as *symbolic learning*. As opposed to *non-symbolic learning*, where, generally, the learned model is a mathematical function (e.g. linear regression, support vector machine or neural network), symbolic models are structures of logical formulas that are directly translatable to natural language. As a matter of fact, symbolic models enclose an explicit knowledge representation, which offers many opportunities: from verifying that the model's thought process is adequate for a given task, to learning new insights by simple inspection of the model, up to allowing the practitioner to manually refine the model at a later time.

While formal logical languages make up a whole research field, in which languages are studied and categorized according to their expressive power and computational properties, symbolic learning has

mainly been studied in the case of two most renowned families of logics: propositional and first-order logics. Propositional symbolic learning includes well-known methods such as *decision trees* and *rule-based models*; because of the low expressive power of propositional logics, these models are easily understandable but can only perform an extremely limited kind of reasoning and can only deal with scalar data. On the other hand, first-order symbolic learning, also known as *inductive logic programming* (ILP), is able to deal with relational data but is only practical within specific, poorly constrained domains; in fact, the higher expressive power of first-order logics comes with a complexity space that is less constrained, ultimately returning the “burden of constraining” to the data itself, which is, therefore, required to specify redundant (and sometimes obvious) logical constraints. Consider, in this perspective, a dataset with spatial components, where each sample consists of an arrangement of objects in a geometrical space (e.g. a dataset of digital images). On the one hand, propositional logic cannot natively deal with such data, as the spatial arrangement of objects cannot be seamlessly squashed into a scalar description. On the other hand, first-order logic is able to capture geometrical relations only when these relations are completely axiomatized, for example, an ILP method can learn patterns depending on the concept of an object *being inside* another but requires the

data to explicitly declare the properties for the relation *being inside* (in this case, irreflexivity, transitivity and asymmetry). However, because relational properties in geometrical spaces arise in a natural way, the last decades have seen a rising theoretical interest toward more constrained logics. Among these, *modal logics* (Blackburn et al., 2001) provide an interesting trade-off between expressive power and complexity; because they are based on implicitly constrained structures, they can express concepts in terms of spatial/temporal entities and relations, but they do not need complex axiomatizations. Moreover, modal logics can be tailored for *temporal* or *spatial* reasoning; thus, extending classical learning schemes to modal logics can lead to models that natively perform temporal/spatial reasoning on time-dependent or spatially distributed data. Similar to what's been done for the case of first-order decision trees (Blockeel & De Raedt, 1998), in recent years, classical decision trees have been extended to modal logics of time and yielded interesting results in time series classification tasks (Coccagna et al., 2022; Manzella et al., 2023; Sciavicco & Stan, 2020).

In this article, we consider an image classification task and show that, when compared to classical decision tree learning, *modal decision tree* learning is able to extract knowledge that is more adequate for the complexity of the task at hand. Specifically, we consider the problem of *land cover classification* (LCC), which is one of many tasks in computer vision where the advent of neural networks induced a change of course. While at the time of writing most of the state of the art in LCC is based on convolutional neural networks (see, e.g. Cao et al., 2018; Santara et al., 2017), a proper literature review reveals how, since the beginning, due to a declared need for explicit classification rules, the task has been frequently addressed using propositional symbolic learning (see, e.g. Goel et al., 2003; Zhang & Wang, 2003). This suggests how, in spite of their benefits in terms of transparency, the inability of known symbolic learning schemes to deal with complex data caused researchers and practitioners to favor higher statistical performances. But, in a way, while the life of black-box models ends with their statistical performances, that of symbolic ones starts with it, and symbolic models enable a continuous interaction between artificial and human intelligence; in this sense, symbolic learning is still a relatively poorly understood field. We perform a systematic experimental evaluation of traditional and spatial decision tree learning methods on five publicly available benchmark datasets for LCC, and finally, we draw some high-level conclusions. Note that, although these datasets are commonly used in literature to evaluate functional methods, none of these methods

addresses interpretability matters, and thus, a comparison against these methods makes little sense; in a similar way, the existing symbolic methods used for the task are generally not spatial in nature (except in Jiang et al., 2012, for which, however, data are not available). As such, the comparison is limited to decision tree learning methods; more specifically, we show how the modal approach displays higher expressive power and performances when compared to the propositional one and how an (explicit) logical theory of the underlying phenomenon can be extracted, discussed, interpreted and translated into natural language, which would not be possible with functional methods. We also leverage the straightforward generalization of decision tree learning methods into *random forest* models (Breiman, 2001), which, despite being less interpretable than single trees, are able to attain performances that are closer to the neural state of the art. Such a generalization can be done both at the propositional and at the spatial level, in a way similar to the study by Manzella et al. (2023). Although they are not immediately comparable to our solution (which is part of a long-term project that aims at generalizing symbolic learning with modal logics in a comprehensive way), there have been a few attempts at logic-based spatial learning. Among them, we mention the study by Malerba et al. (2005), in which the authors extract a regression function that depends on the spatial relationships among objects in an image, and it is a more database-oriented work than a pure learning method, and the study by Dubba et al. (2015), in which a customized ILP-like approach is used to extract knowledge from video samples. These approaches are mildly similar to ours, but they work with first-order logics and thus require first-order descriptions of the data before the learning can take place.

Related work

Land cover classification

Land cover classification (LCC) typically refers to the task of classifying pixels in remotely sensed images, associating each pixel with a label that captures the use of the land it depicts (e.g. *forest*, *urban area* or *crop field*). Despite being an instance of standard *image segmentation*, LCC has a few peculiarities that allow it to be dealt with as an *image classification* task. Images of this kind are usually captured from satellite or aerial sensors, and due to the altitude from which the images are captured, a single pixel carries the average electromagnetic response of a wide surface area (e.g. one to hundreds of square meters); as such, the classification of a pixel in the image typically depends only on a close neighborhood of pixels around it. Additionally, the imagery involved is often

hyperspectral, that is, composed of many spectral channels, describing the strength of signals at different electromagnetic wavelengths; thus, each pixel holds a large number of values (e.g. usually in the hundreds) that collectively describe the nature of its content satisfactorily. Because of these key features, the task can be dealt with using a *per-pixel* approach: first, a dataset of pixel samples, typically represented with feature vectors, is extracted from the image; then, a machine learning method is trained on the pixel samples; finally, in the operational phase, each pixel is classified independently. Although not related to this work, it should be noted that a plethora of *object-based* approaches exists as well; refer to Kucharczyk et al. (2020) for a recent review on the topic.

LCC has a long history of being tackled with per-pixel symbolic learning, starting from the most influential articles, dating back to 1995–2003 (Friedl & Brodley, 1997; Goel et al., 2003; Kartikeyan et al., 1995; Pal & Mather, 2003), and continuing to more recent works (Belgiu et al., 2014; Berhane et al., 2018). All the work in the symbolic realm is based on propositional logics; most of it focuses on decision trees (Berhane et al., 2018; Friedl & Brodley, 1997; Goel et al., 2003; Kulkarni & Shrestha, 2017; Monteiro & Murphy, 2011; Pal & Mather, 2003; Phiri et al., 2020; Xu et al., 2005), with only a few examples of rule-based models (Belgiu et al., 2014; Zhang & Wang, 2003). Although some authors made remarks on how functional models can be more statistically accurate (Goel et al., 2003; Phiri et al., 2020), all the literature on this side agrees on the importance of having clear classification rules for the task, as they can be: a) interpreted in order to identify spectral similarities and differences between different classes and b) validated and used for defining *standards* for class membership conditions. Nonetheless, the per-pixel LCC literature since 2015 has been flooded with highly cited works using neural network-based methods, effectively raising the bar in terms of statistical accuracy (Audebert et al., 2019; Cao et al., 2018; Hong et al., 2022; Hu et al., 2015; Jiang et al., 2019; Lee & Kwon, 2017; Li et al., 2017; Mou et al., 2017; Roy et al., 2020; Santara et al., 2017). One of the reasons why these approaches (mainly based on convolutional networks, but also recurrent networks (Audebert et al., 2019; Mou et al., 2017) and transformer models (Hong et al., 2022)) are more effective than symbolic methods is that they leverage intrinsic spatial localities that occur in the data. In fact, while the vast majority of symbolic work relies on propositional logic, and a simple feature vector consisting only of the pixel's own spectral footprint (thus disregarding the spatial structure of the image), these deep neural networks can factor in a close neighborhood of pixels around the pixel to be classified,

ultimately capturing inter-pixel dependencies and correlations.

The body of symbolic work accounting for inter-pixel correlations is likely confined by the inability of propositional methods to handle spatial data; to our knowledge, it is limited to (Jiang et al., 2012), which proved the efficacy of integrating the (local) feature vector with additional scalar features capturing information about the neighboring pixels. This approach is, again, based on propositional trees and relies on a functional feature extraction step, which may undermine the interpretability of the model (on which no comments are made throughout the article).

Formalisms for qualitative spatial reasoning

Qualitative spatial reasoning (QSR) has many important real-world applications, and there is active research on deriving formal systems for reasoning about space. Throughout the last 30 years, many formalizations have been proposed for different purposes, mostly in the form of *binary spatial calculi* but sometimes also in terms of formal logics (Renz & Nebel, 2007). Two major structural choices for deriving a spatial logic concern the definitions of a set of *spatial entities* and a set of *binary relations* between the entities. Depending on the application, objects are classically chosen to be represented as points or regions of a specific kind (e.g. convex regions). Points are elementary entities that cannot express the extension of objects, thus they are not suitable in contexts where objects have diverse shapes or when mereological aspects of space are relevant (e.g. grade of overlap between regions). Regions, on the other hand, can express extension but are more complex to deal with. When regions are considered, they are usually assumed to be *connected*, which is in agreement with an implicit locality hypothesis on the objects of the reasoning. Although limited in terms of expressive power, convex regions are typically preferred as extended entities, or they are used to approximate complex shapes (e.g. by means of *convex hulls* Cohn, 1995). Bidimensional convex regions, providing good trade-offs between expressive power and complexity, include *triangles* and *convex quadrilaterals*. Regions can be further constrained to achieve higher specificity or lower complexity, for example, forcing all entities to be of the same size can be beneficial for specific domains such as *optical character recognition*. Another sensible constraint is *orthogonality*. In the case of polygons, forcing the edges to be parallel to a given reference set of axes limits the set of possible relations and sometimes lowers computational costs; in particular, rectangles with edges parallel to the axes

(such as in Rectangle Algebra, presented in Balbiani et al., 1998) have been deeply used in QSR due to their computational benefits.

As for relation sets, different aspects can be considered (e.g. distance, size and shape of the entities), and two of the most studied relational aspects are *directionality* and *topology*. Directional relations account for the relative spatial arrangement of the two objects (e.g. *next to* and *to the right of*); they are not invariant under rotation/reflection and are not easily definable for generic types of regions. On the algebraic side, many calculi have been proposed, both with punctual entities (Frank, 1996) and regions of various types (Balbiani et al., 1998; Navarrete et al., 2013; Skiadopoulos & Koubarakis, 2004; Skiadopoulos et al., 2007). Directional algebras could, in principle, give rise to their modal logic counterparts, but only a few attempts have been made in this sense (Bresolin et al., 2010; Marx & Reynolds, 1999; Montanari et al., 2009; Morales Nicolás et al., 2007). While in this context an absolute frame of reference is generally adopted, there are examples of spatial logics and calculi that account for the *orientation* of objects and are able to describe relations such as *in front of*, *behind* or *facing each other* (Freksa, 1992; Walega & Zawidzki, 2019). As for topological relations, they generally address the different modalities of intersection between regions and their boundaries, which makes them invariant under rotation and reflection; being purely qualitative and easily definable for generic regions, topological approaches are often preferred to directional ones. The most popular characterization of topological binary relations is the set of Region Connection Calculus relations, often referred to as RCC8 (Egenhofer et al., 1993; Randell et al., 1992). In its extended form, the set includes eight relations: *disconnected*, *externally connected*, *partially overlapping*, *tangential proper part*, *non-tangential proper part*, *tangential proper part inverse*, *non-tangential proper part inverse*, and *identity*. Coarser sets can be derived by considering unions of RCC8 relations, deriving, for example, RCC5. In the study by Lutz and Wolter (2006), the modal logics that are entailed by these sets of relations, namely \mathcal{L}_{RCC8} and \mathcal{L}_{RCC5} , are studied.

It is of note that, compared to topological algebras, *jointly exhaustive* region-based directional algebras (namely, algebras in which each pair of entities is in at least a relation) tend to have a higher number of relations, as well as a higher granularity; moreover, their relations can generally be partitioned and joined to capture topological aspects and, ultimately, to form coarser topological systems. With this in mind, and also considering the convenient properties of orthogonal rectangles, we see Rectangle Algebra (Balbiani et al., 1998), with its 169 directional relations (see Figure 2), as the most noteworthy formalization for general-purpose, bidimensional QSR; this observation is also in line with

the renowned importance of the 13 (one-dimensional) relations (Allen, 1983) onto which Rectangle Algebra is based on.

Spatial decision tree learning

While the behavior of classical decision trees based on propositional logic is well-defined for the case of tabular data, they provide no native way for dealing with image data; in fact, it is often the case that, when dealing with images, machine learning practitioners apply feature engineering methods that preventively reduce the data to a scalar description of itself, prior to training a decision tree. We argue that this process denatures the structure of data, and that it can easily challenge the transparency of the model. Instead, we propose a solution along the lines of what has been done for the case of temporal data (Coccagna et al., 2022; Manzella et al., 2023; Sciavicco & Stan, 2020), where the well-known interval temporal logic \mathcal{HS} (Halpern & Shoham, 1991) has been used to allow a decision tree to perform temporal modal reasoning. In this section, we give a formal presentation of Rectangle Logic (or \mathcal{HS}^2 , logical counterpart to Rectangle Algebra), which is capable of express directional spatial patterns, such as *there exists a rectangular region in the image with a high level of red, to the left of another rectangular region with a low level of green*, as well as topological spatial patterns involving RCC relations.

\mathcal{HS}^2 A spatial modal logic of rectangles

Inspired by both the interval temporal logic \mathcal{HS} (Halpern & Shoham, 1991) and Rectangle Algebra (Balbiani et al., 1998), we hereby introduce the spatial modal logic \mathcal{HS}^2 based on orthogonal rectangles on a finite, bidimensional space. \mathcal{HS} is based on the set of 13 binary relations that arise from a natural definition of intervals on a linear order (Allen, 1983); using the original notation from (Halpern & Shoham, 1991), we depict their informal semantics in Figure 1 and refer to them as follows:

$$\mathcal{X} = \{A, L, B, E, D, O, \bar{A}, \bar{L}, \bar{B}, \bar{E}, \bar{D}, \bar{O}, =\}.$$

Now, let

$$\begin{aligned} \mathbb{D}_1 &= \langle D_1, < \rangle = [1, \dots, N_1] \\ \mathbb{D}_2 &= \langle D_2, < \rangle = [1, \dots, N_2] \end{aligned}$$

be two finite, linearly ordered sets, and let $\mathbb{D}^2 = \mathbb{D}_1 \times \mathbb{D}_2$ be a finite and discrete geometrical space. Elements of \mathbb{D}^2 , called *points*, are denoted as $\pi = (x_1, x_2)$. In analogy with interval temporal logic, a *rectangle* in \mathbb{D}^2 is an object of the type:

$$[(x_1, y_1), (x_2, y_2)],$$

Name	Definition w.r.t. an interval structure	Example
A (after)	$[x, y]R_A[w, z] \Leftrightarrow y = w$	
L (later)	$[x, y]R_L[w, z] \Leftrightarrow y < w$	
B (begins)	$[x, y]R_B[w, z] \Leftrightarrow x = w \wedge z < y$	
E (ends)	$[x, y]R_E[w, z] \Leftrightarrow y = z \wedge x < w$	
D (during)	$[x, y]R_D[w, z] \Leftrightarrow x < w \wedge z < y$	
O (overlaps)	$[x, y]R_O[w, z] \Leftrightarrow x < w < y < z$	

Figure 1. Representation of 6 of the 13 Allen's relations. The seven remaining relations are the inverses of the six relations depicted, and the identity relation $R_=_$. However, note that $R_=_$ is not logically interesting, and, in fact, \mathcal{HS} only encompasses 12 modal operators.

with $1 \leq x_1 < y_1 \leq N_1$, and $1 \leq x_2 < y_2 \leq N_2$. From a geometrical point of view, rectangles are essentially the combination of two intervals: one along the horizontal axis and the other along the vertical axis; as such, the notion of π *belonging* to a rectangle ($\pi \in [(x_1, y_1), (x_2, y_2)]$) can be defined using the order relations $<$, and different such notions can be defined depending on the particular application. Now, let $\mathbb{K}(\mathbb{D}^2)$ be the set of all rectangles that can be formed on \mathbb{D}^2 ; in the following, we use r, s, \dots to denote rectangles. Because there are 13 distinct relations between any two intervals in a linear order, there are $13^2 = 169$ relations between any two rectangles (Figure 2). In a single dimension, we denote relations as R_X , where $X \in \mathcal{X}$, $R_=_$ denotes the identity relation, and $R_{\bar{X}}$ denotes the inverse of R_X . In the bidimensional version, any relation can be seen as a tuple of two one-dimensional relations and can be

denoted by R_{X_1, X_2} , with $X_1, X_2 \in \mathcal{X}$. Lifting such relations to the modal level entails introducing $169 - 1 = 168$ unary modal operators of the type:

$$\langle X_1, X_2 \rangle,$$

where $X_1, X_2 \in \mathcal{X}$, with the additional constraint that X_1 and X_2 cannot be both the identity relation (in fact, because identity is not a logically interesting relation, $\langle R_=_, R_=_ \rangle$ would have no role in this context). Formulas of Rectangle Logic are, therefore, obtained as follows:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \langle X_1, X_2 \rangle \varphi,$$

where p is a propositional letter from a given set \mathcal{P} . Such formulas are easily interpreted in a *bi-bidimensional spatial model* $\mathcal{M} = \langle \mathbb{K}(\mathbb{D}^2), EV \rangle$, where EV is a valuation function defined as follows:

$$EV : \mathbb{K}(\mathbb{D}^2) \mapsto 2^{\mathcal{P}},$$

	\bar{L}	\bar{A}	\bar{O}	\bar{E}	\bar{D}	B	$=$	\bar{B}	D	E	O	A	L
\bar{L}													
\bar{A}													
\bar{O}													
\bar{E}													
\bar{D}													
B													
$=$													
\bar{B}													
D													
E													
O													
A													
L													

Figure 2. Pictorial examples for the 169 relations of rectangle algebra (Balbiani et al., 1998). Each relation is the Cartesian product of two one-dimensional interval relations (Allen, 1983); therefore, rectangle relations can be naturally arranged in a matricial structure. For each relation X , a reference entity r_1 (grey) and a secondary entity r_2 (transparent), with $(r_1, r_2) \in X$.

which assigns to each rectangle the set of all and only propositional letters that are true on it, by means of the following truth relation:

$$\begin{array}{lll} \mathcal{M}, r \models p & \text{iff} & r \in EV(p) \\ \mathcal{M}, r \models \neg\varphi & \text{iff} & \mathcal{M}, r \not\models \varphi \\ \mathcal{M}, r \models \varphi \vee \psi & \text{iff} & \mathcal{M}, r \models \varphi \text{ or } \mathcal{M}, r \models \psi \\ \mathcal{M}, r \models \langle X_1, X_2 \rangle \varphi & \text{iff} & \exists s \text{ s.t. } (r, s) \in R_{X_1, X_2}, \mathcal{M}, s \models \varphi. \end{array}$$

The corresponding universal operator $[X_1, X_2]$ of an existential modality $\langle X_1, X_2 \rangle$ is defined in the standard way:

$$[X_1, X_2]\varphi = \neg\langle X_1, X_2 \rangle\neg\varphi.$$

While encompassing as many as 168 different modal operators, \mathcal{HS}^2 is intuitive. Its relations are mutually exclusive and jointly exhaustive with respect to $\mathbb{K}(\mathbb{D}^2)$; as such, exactly one relation holds for each pair of rectangles in $\mathbb{K}(\mathbb{D}^2)$, and the global existential operator, that allows to express global patterns (i.e. “there exists a rectangular region *anywhere* ...”), can be defined by disjunction of all modal operators:

$$\langle G \rangle \varphi = \varphi \vee \bigvee_{X_1, X_2} \langle X_1, X_2 \rangle \varphi,$$

with X_1 and X_2 satisfying the same constraints above. In a similar way, disjunctions of \mathcal{HS}^2 operators can give rise to operators for RCC8 and RCC5 through derived relations (and thus, derived modal operators) achieved via union (i.e. disjunction) of suitable subsets of the 168 relations; for example, the topological relation *disconnected*, when interpreted on $\mathbb{K}(\mathbb{D}^2)$, can be defined as follows:

$$R_{DC} = \bigcup_{X \in \{A, L, B, E, D, O, =\}} (R_{\bar{L}, X} \cup R_{L, X} \cup R_{X, \bar{L}} \cup R_{X, L}).$$

Thus, one can easily define two topological rectangle logics: \mathcal{HS}_{RCC8}^2 (seven relations, shown in Figure 3a) and \mathcal{HS}_{RCC5}^2 (four relations, shown in Figure 3b); these fragments of \mathcal{HS}^2 correspond to Lutz and Wolter’s \mathcal{L}_{RCC8} and \mathcal{L}_{RCC5} , when interpreted on orthogonal rectangles.

Although \mathcal{HS}^2 has not been studied per se, the literature that concerns its fragments is very wide. Very briefly, it is worth recalling that satisfiability for \mathcal{HS} , its one-dimensional version, is undecidable (Halpern & Shoham, 1991), and that various strategies have been considered in the literature to define fragments or variants of \mathcal{HS} with better computational behavior. These include constraining the underlying temporal structure (Montanari et al., 2002), restricting the set of modal operators (Aceto et al., 2016; Bresolin et al., 2014), softening the semantics to a reflexive one (Marcinkowski & Michaliszyn, 2014; Montanari et al., 2010), restricting the nesting of modal operators (Bresolin et al., 2014), restricting the propositional power of the languages (Bresolin et al., 2017) and considering *coarser* interval temporal logics based on

interval relations that describe a less precise relationship between intervals (similar to what topological relations do) (Muñoz-Velasco et al., 2019). As for \mathcal{HS}^2 , only the topological fragments for \mathcal{HS}_{RCC8}^2 and \mathcal{HS}_{RCC5}^2 have been studied by Lutz and Wolter (2006), and their satisfiability problem is undecidable as well, even under very simple assumptions (e.g. for every class of linear orders on which \mathbb{D}_1 and \mathbb{D}_2 are based), or can be proven so by exploiting the results on the one-dimensional case. In general, one can expect deduction in \mathcal{HS}^2 to be a computationally hard problem even under very restrictive assumptions, such as finite domains. Although, in the spirit of the existing work for interval temporal logic, one can imagine studying fragments of \mathcal{HS}^2 , in this machine learning context, we only focus on inductive aspects, for which expressive power is a more important quality. With this regard, different fragments of \mathcal{HS}^2 (e.g., \mathcal{HS}_{RCC8}^2) express different aspects in the spatial arrangement of the entities, and the fragment to be used in a given context remains a hyperparameter of the learning algorithm.

A theory of modal decision trees for image classification

Let an *image dataset* be a set $\mathcal{I} = \{I_1, \dots, I_m\}$ of m images over a variable set \mathcal{V} . For the sake of simplicity, we assume that every image has the same size $W \times H$, thus each image is a three-dimensional tensor $I_i \in \mathbb{R}^{W \times H \times |\mathcal{V}|}$, holding the spatial distribution of variables in \mathcal{V} , across two axes of sizes W (width) and H (height), respectively. From a logical perspective, images can be interpreted as models of \mathcal{HS}^2 in a geometrical space $\mathbb{D}^2 = \mathbb{D}_1 \times \mathbb{D}_2$, with $\mathbb{D}_1 = [1, \dots, W + 1]$, $\mathbb{D}_2 = [1, \dots, H + 1]$, by means of the following point-to-rectangle membership definition:

$$(z, t) \in [(x_1, y_1), (x_2, y_2)] \text{ iff } x_1 \leq z < y_1 \wedge x_2 \leq t < y_2.$$

This definition slightly differs from the standard definition of membership used across the logic community; however, unlike the standard one, this interpretation allows us to associate each rectangle in $\mathbb{K}(\mathbb{D}^2)$ with a rectangular subregion of $w \times h$ pixels with $h, w \geq 1$ (e.g. we can capture 1×1 pixel regions with rectangles r composed of two unit intervals, i.e. $r = [(x_1, x_1 + 1), (x_2, x_2 + 1)]$). A *classification image dataset* is a dataset associated with a *class variable* with a categorical domain $\mathcal{C} = \{C_1, \dots, C_l\}$, i.e. each image I_i is labeled with a *class value* $C_i \in \mathcal{C}$ (also referred to as the *ground truth*). The *class* of an image $I \in \mathcal{I}$ is denoted by $C(I)$. Although there are cases where variables have different interpretations, variables in digital

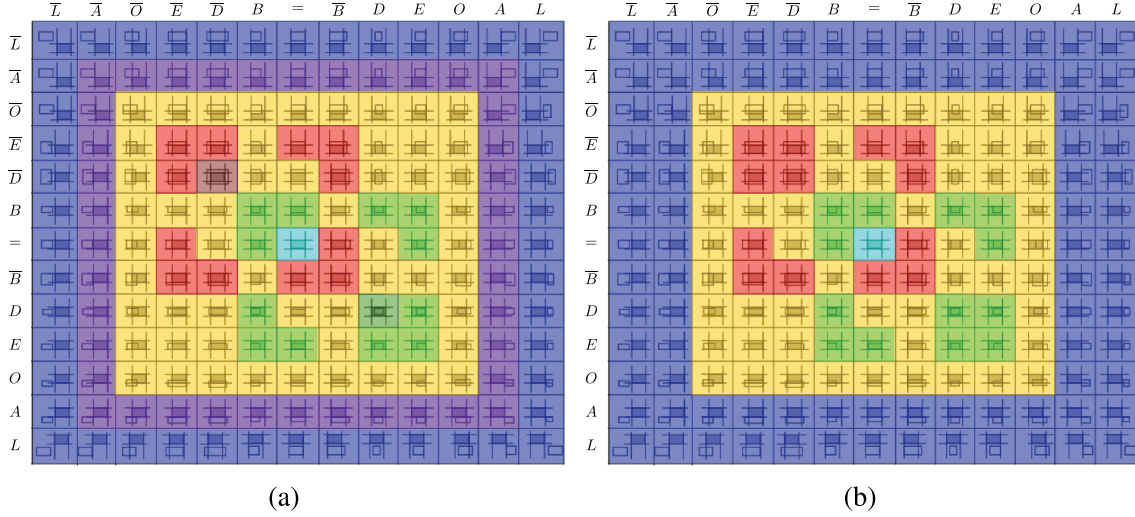


Figure 3. Partitions of HS^2 relations into HS^2_{RCC8} (a) and HS^2_{RCC5} (b) relations. HS^2_{RCC8} relations are: *disconnected* (blue), *externally connected* (violet), *partially overlapping* (yellow), *tangential proper part* (light green), *non-tangential proper part* (dark green), *tangential proper part inverse* (red), *non-tangential proper part inverse* (maroon), and *identity* (turquoise). HS^2_{RCC5} relations are: *discrete from* (blue), *partially overlapping* (yellow), *proper part* (green), *proper part inverse* (red) and *identity* (turquoise).

images often represent the distribution of electromagnetic radiation at different wavelengths. In this case, images usually cover the visible spectrum and encompass only a few variables (e.g. three in the standard RGB configuration), but there are also cases of variables outside the visible spectrum (*hyperspectral imagery*). For a spatial variable $V \in \mathcal{V}$, let $V(\pi)$ be the value of V at point π ; note that $V((z, t))$ is not defined when $z = W + 1$, or when $t = H + 1$.

We shall describe the properties of images by expressing scalar conditions on the rectangles. Such properties are called *propositional split-decisions* (or, more simply, *propositional decisions*), and they match the set of propositional letters:

$$\mathcal{S}_{prop} = \mathcal{P} = \{f(V) \bowtie v\},$$

where $\bowtie \in \{>, \geq, <, \leq\}$ is a *test operator*, f is a *feature function* and v a real number. In the most general case, properties computed through feature functions can involve many spatial variables; for example, a feature function can consist of a deep neural network for object recognition. However, in this formulation, f is a scalar descriptor for a single variable V within the given rectangle. Single-variable rectangle feature functions that are especially interesting for interpretability purposes are the minimum value function \min , the maximum value function \max , and their softened versions \min_γ and \max_γ , with $\gamma \in (0, 1]$, computing something quite similar to the quantiles of the variables values). In particular, we note that $\min(V) > v$ (resp., $\max(V) < v$) can be interpreted as “within the rectangle, variable V is always greater (resp., less) than v ”, and that $\min_\gamma(V) > v$

(resp., $\max_\gamma(V) < v$) can be interpreted as “within the rectangle, at least γ of the points have a value for variable V that is greater (resp, less) than v ”. Fixed an image $I \in \mathcal{I}$, for a proposition to hold on a rectangle $r = [(x_1, y_1), (x_2, y_2)]$ is denoted as:

$$I, r \models f(V) \bowtie v.$$

The presented theory of spatial, modal decision trees is based on \mathcal{HS}^2 (although, as mentioned above, the actual fragment used is a hyperparameter), thus, on top of propositional decisions, the language of spatial decision trees encompasses a set of *modal decisions*:

$$\mathcal{S}_{mod} = \{\langle X_1, X_2 \rangle f(V) \bowtie v\},$$

where $\langle X_1, X_2 \rangle$ is a \mathcal{HS}^2 modality. Together, propositional and modal decisions form a set of *decisions*:

$$\mathcal{S} = \mathcal{S}_{prop} \cup \mathcal{S}_{mod}.$$

A binary *spatial decision tree* τ is a structure:

$$\tau = (T, E, \rho, l, s),$$

where (T, E) is a directed binary tree rooted in $\rho \in T$, $T^i = T \setminus T^\ell$ the subset of its *internal nodes*, with any internal node v having a *left* (resp. *right*) *child* $\ast^r(v)$ (resp., $\ast^l(v)$), $T^\ell \subseteq T$ is the subset of *leaf nodes*, $l : T^\ell \rightarrow \mathcal{C}$ is a *leaf-labelling function*, and $s : T^i \rightarrow \mathcal{S}$ is an *internal node-labelling function*. A graphical representation of an example tree can be seen in Figure 4. Although each internal node is associated with a split decision, it is common to depict the split decision of a given node along its outgoing left edge and its negation along the right outgoing edge. In a similar way, it is convenient to introduce the edge-labelling function; let $e = (v_0, v_1) \in E$, the function is defined as follows:

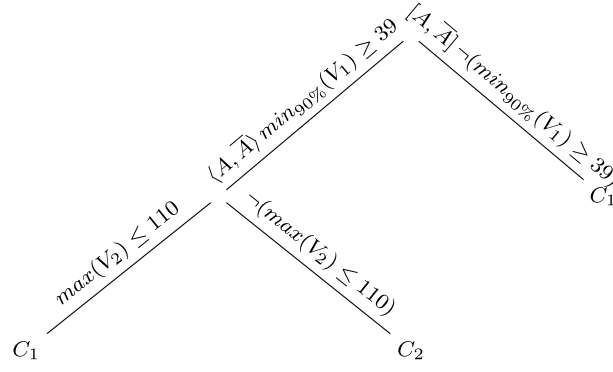


Figure 4. Example of a spatial decision tree with a spatial decision at the root node, followed by a propositional decision at the left child.

$$s'(e) = \begin{cases} s(v_0); & \text{if } v_1 = v_0 \\ \neg p & \text{if } v_1 = v_0 \text{ and } s(v_0) = p; \\ [X_1, X_2] \neg p & \text{if } v_1 = v_0 \text{ and } s(v_0) = \langle X_1, X_2 \rangle p. \end{cases}$$

To define the semantics of spatial decision trees on an image dataset, we first need to define how an image is *evaluated* (in this case, *classified*) by a tree, and thus how decisions are tested. With traditional trees based on propositional logic, each decision can be directly evaluated on a sample, and each internal node simply redirects the sample to one of its children, according to whether the decision holds on the sample, until a leaf is reached. With spatial decision trees, decisions must be relativized to rectangles, and some form of memory can be introduced in the operational classification policy; we address this need by introducing the concept of *witness for a decision*, i.e. a rectangle where the decision holds, and by propagating a set of witnesses through the evaluation procedure. An image $I \in \mathcal{I}$, then, starts at the root node ρ , with an initial non-empty set of rectangles $K_0 \in \mathbb{K}(\mathbb{D}^2)$ and is classified as $\tau(\rho, I, K_0)$ by the following recursive definition:

$$\tau(v, I, K) = \begin{cases} l(v) & \text{if } v \in T^\ell; \\ \tau((v), I, w(I, K, S)) & \text{if } \exists r \in K \text{ s.t. } I, r \models S; \\ \tau((v), I, K) & \text{if } \nexists r \in K \text{ s.t. } I, r \models S. \end{cases}$$

In the case that S holds on at least one of the current witnesses, w computes a new witness set for S and allows the tree to visit different areas of the image, ultimately checking for the existence of modal spatial patterns. Although the behavior is quite similar for the cases of propositional and spatial decisions, we conveniently differentiate the two cases for defining w . If $S = \langle X_1, X_2 \rangle (f(V) \bowtie v)$, then $w(I, K, s)$ is defined as follows:

$$\{s \in \mathbb{K}(\mathbb{D}^2) \mid \exists r \in K \text{ s.t. } (r, s) \in R_{X_1, X_2}, I, s \models f(V) \bowtie v\},$$

while if $S = f(V) \bowtie v$, then it is defined as:

$$\{s \in K \mid I, s \models f(V) \bowtie v\}.$$

Observe that K_0 represents the initial condition for spatial decision tree evaluation, and it can become a parameter of the tree model. In the most general case, one may want K_0 to depend on the image to

evaluate; for example, it could consist of a single rectangle (interpreted as a privileged observation point) as in $K_0^I = \{r_0^I\}$, where r_0^I is, for example: a) the unit rectangle at the center of the image: $[(\frac{W+1}{2}, \frac{W+1}{2} + 1), (\frac{H+1}{2}, \frac{H+1}{2} + 1)]$, or the largest rectangle, consisting of the image frame: $[(0, W + 1), (0, H + 1)]$. Alternatively, with the suitable fragment, global patterns can be captured by fixing any initial rectangle as r_0 and allowing decisions of the kind $\langle G \rangle f(V) \bowtie v$. Similarl to the case of traditional decision trees, any spatial tree provides a classification rule for each leaf, but the rule associated to a given path requires a more complex formalization. We associate any path $\pi = v_0 v_1 \dots v_h$ with its *path-formula* φ_π , inductively defined as follows. If $h = 0$, then $\varphi_\pi = T$; if $h = 1$, then $\varphi_\pi = s'(e)$; finally, $h > 1$, let $\lambda = s'(e)$ and $\pi' = v_1 v_h$, then φ_π is:

$$\varphi_\pi = \begin{cases} (\lambda \wedge \varphi_{\pi'}) & \text{if } \lambda = p \text{ and } v_2 = (v_1) \\ \langle X \rangle (p \wedge \varphi_{\pi'}) & \text{if } \lambda = \langle X \rangle p \text{ and } v_2 = (v_1) \\ \lambda \wedge (\lambda \rightarrow \varphi_{\pi'}) & \text{if } \lambda = p \text{ and } v_2 = (v_1) \\ \lambda (p \rightarrow \varphi_{\pi'}) & \text{if } \lambda = \langle X \rangle p \text{ and } v_2 = (v_1) \\ (\lambda \wedge \varphi_{\pi'}) & \text{otherwise} \end{cases}$$

where $e = (v_0, v_1)$ and $\pi' = v_1 v_h$. Finally, for each leaf node $\ell \in T^\ell$, the rule associated to the leaf is an object $\varphi_\ell \Rightarrow l(\ell)$, where the *leaf-formula* φ_ℓ is defined as follows:

$$\varphi_\ell = \bigwedge_{\pi \in \text{prefix}(\rho\ell)} \varphi_\pi.$$

Considering the operational semantics of decision trees (i.e. how the evaluation occurs), any leaf-formula φ_ℓ can be thought as a sufficient condition for class $l(\ell)$. To fix the ideas, consider, again, the tree in [Figure 4](#); given an initial rectangle r_0 , this tree comprehends the following sufficient condition for class C_2 (i.e. the leaf-formula of the only leaf labelled with C_2):

$$T \wedge \langle A, \overline{A} \rangle p \wedge \langle A, \overline{A} \rangle p \wedge [A, \overline{A}] (p \rightarrow q),$$

where $p = \min_{90\%}(V_1) \geq 39, q = \max(V_2) \leq 110$. Note that this formula can be simplified into $\langle A, \bar{A} \rangle p \wedge [A, \bar{A}] (p \rightarrow q)$; therefore, the leaf for C_2 provides the following classification rule, as translated to natural language: *IF there exists at least one rectangle northwest of r_0 , touching the top-right corner of r_0 , where $\min_{90\%}(V_1) \geq 39$, but all such rectangles also have $\max(V_2) > 110$, THEN C_2* . In this case, where the model comprehends a unique leaf labelled with C_2 , the aforementioned leaf formula is also a necessary condition for this class. In general, however, a given class C_i will appear in many leaves; thus, one can define its *class-formula* by joining all the leaf-formulas with leaves labelled with C_i , as in:

$$\varphi_{C_i} = \bigvee_{\ell \in T^\ell, l(\ell)=C_i} \varphi_\ell.$$

The class-formula for C_i makes a necessary condition for C_i ; thus, we write $\varphi_{C_i} \Leftrightarrow C_i$.

Given an image dataset \mathcal{I} , any classification rule for a class C_i can be evaluated considering the number of instances n that reached to the corresponding leaf, the number of instances c that the rule correctly classifies, the number of available instances $t = |\mathcal{I}|$, and the number of instances that belong to C_i , denoted here as t_{C_i} . Typical performance metrics for rule extraction are *support*, *confidence*, *lift* and *conviction*, which are defined as follows:

$$\begin{aligned} \text{support} &= \frac{n}{t} \\ \text{confidence} &= \frac{c}{n} \\ \text{lift} &= \frac{\frac{c}{n}}{\frac{t_{C_i}}{t}} \\ \text{conviction} &= \frac{1 - \frac{t_{C_i}}{t}}{1 - \frac{c}{n}}. \end{aligned}$$

It is important to point out that the very essence of the interpretability of decision trees lies in the possibility of deriving and evaluating such rules, translating them to natural language and manipulating them. It should be noted that, similar to any other computational model, decision tree models and their associated rules can be large and complex; this mainly depends on the phenomenon that a given tree model encloses. However, split decisions, leaf formulas and class formulas are forms of explicit knowledge that can always be translated to natural language and thus conveyed to human intelligence. Ultimately, it is convenient to lift the sample-level operational semantics to the case of multiple images. Images in a dataset can be evaluated in bulk by finally introducing the concept of *dataset split*. With \mathcal{K} denoting a set of witness sets (i.e. a witness set for each image), let \mathcal{I}_v and \mathcal{K}_v be the *dataset and witnesses associated to a node v* ;

when evaluating a dataset \mathcal{I} with initial conditions \mathcal{K} , these two sets are, first, associated to the root node: $\mathcal{I}_\rho := \mathcal{I}, \mathcal{K}_\rho := \mathcal{K}$. Then, a dataset split happens at any internal node v when evaluating the local split-decision $s(v)$ on each image of \mathcal{I}_v , considering \mathcal{K}_v , i.e. the witness set associated to each image at the node. The split partitions \mathcal{I}_v (resp., \mathcal{K}_v) into the subset $\iota_{\mathcal{K}^s}(v)$ (resp., $\kappa_{\mathcal{K}^s}(v)$) of images (resp., witness sets) that are redirected to the left child, and the subset $\iota_{\mathcal{K}^a}(v)$ (resp., $\kappa_{\mathcal{K}^a}(v)$) of those that are redirected to the right child. Evaluating a tree τ on a dataset \mathcal{I} entails classifying all $I \in \mathcal{I}$ and comparing the resulting class with the ground truth; such a comparison gives rise to a confusion matrix, from which performance metrics such as overall accuracy, mean recall and κ coefficient can be computed.

Entropy-based learning of spatial decision trees

The decision tree model was originally defined by (Quinlan, 1986), which resulted in the development of the learning algorithms C4.5 (Quinlan, 1993), which includes the possibility of dealing with numerical variables as well. *Spatial C4.5* extends C4.5 to deal with image datasets, preserving its driving principles. Because learning optimal trees is a NP-hard problem at the propositional level already (Hyafil & Rivest, 1976), so is at the spatial level, and *Spatial C4.5*, just like C4.5, implements a suboptimal, greedy approach generally known as *entropy-based learning*. Let ξ_i be the fraction of images labeled with class C_i in a dataset \mathcal{I} with ℓ distinct classes. Then, the *information conveyed* by \mathcal{I} (or *entropy* of \mathcal{I}) is computed as follows:

$$\text{Info}(\mathcal{I}) = - \sum_{i=1}^{\ell} \xi_i \log \xi_i.$$

Intuitively, the entropy is inversely proportional to the purity degree of \mathcal{I} with respect to the class values. Let τ be an interval temporal decision tree, v be an internal node and \mathcal{I}_v be the dataset associated to v . Then, the *information conveyed* by the split happening in v is:

$$\text{InfoSplit}(\tau, v) = \frac{|\mathcal{I}_{\mathcal{K}^s}(v)|}{|\mathcal{I}_v|} \cdot \text{Info}(\mathcal{I}_{\mathcal{K}^s}(v)) + \frac{|\mathcal{I}_{\mathcal{K}^a}(v)|}{|\mathcal{I}_v|} \cdot \text{Info}(\mathcal{I}_{\mathcal{K}^a}(v)).$$

Now, the *entropy gain* of the split can be defined as follows:

$$\text{InfoGain}(\tau, v) = \text{Info}(\mathcal{I}_v) - \text{InfoSplit}(\tau, v).$$

Algorithm 1: High-level description of *Spatial C4.5*.

```

input :  $\mathcal{I}$  – image dataset
output:  $\tau$  – spatial decision tree
function SpatialC4.5( $\mathcal{I}, \mathcal{K}$ ):
    Preprocess( $\mathcal{I}$ )
     $\tau \leftarrow \text{Learn}(\mathcal{I}, \mathcal{K})$ 
    return  $\tau$ 
end
input :  $\mathcal{I}$  – image dataset
output:  $\nu$  – decision node
function Learn( $\mathcal{I}, \mathcal{K}$ ):
    if some stopping condition applies then
        return CreateLeafNode( $\mathcal{I}$ )
     $S \leftarrow \text{FindBestSplitDecision}(\mathcal{I}, \mathcal{K})$ 
     $(\mathcal{I}_{\mathcal{L}'(\nu)}, \mathcal{K}_{\mathcal{L}'(\nu)}, \mathcal{I}_{\mathcal{R}(\nu)}, \mathcal{K}_{\mathcal{R}(\nu)}) \leftarrow \text{Split}(\mathcal{I}, S)$ 
     $\nu \leftarrow \text{CreateNode}(\mathcal{I}, \mathcal{K})$ 
     $\nu.\text{left} \leftarrow \text{Learn}(\mathcal{I}_{\mathcal{L}'(\nu)}, \mathcal{K}_{\mathcal{L}'(\nu)})$ 
     $\nu.\text{right} \leftarrow \text{Learn}(\mathcal{I}_{\mathcal{R}(\nu)}, \mathcal{K}_{\mathcal{R}(\nu)})$ 
    return  $\nu$ 
end

```

Given an image dataset \mathcal{I} , the language of all possible decisions, and therefore of all possible splits, is implicitly defined, and the generic, high-level *Spatial C4.5* algorithm works as shown in Alg. 4. This learning algorithm returns a binary tree τ , in which every node is associated with a subset of \mathcal{I} , and every leaf node is associated with a class that depends on the labeling policy in use. The typical policy involves labeling a leaf with the class that occurs the most in the associated dataset. As in the propositional case, the stopping conditions, also called *pre-pruning* conditions, depend on the particular implementation and are subject to statistical evaluations. Typical conditions include testing the purity of the dataset associated with the candidate-leaf node, its cardinality or the information gain that may emerge from a further splitting step.

At both the propositional and spatial levels, the theoretical complexity of C4.5 is polynomial; however, at the spatial level, the space of all possible decisions is much generally larger (especially given the large number of relations), and the cost of testing each decision, besides being polynomial with respect to the cardinality of the dataset, is also polynomial with respect to the number of rectangles, namely $|\mathbb{K}(\mathbb{D}^2)|$, which can easily explode, even with small images. The parameters of *Spatial C4.5* include all classical propositional parameters (e.g. the stopping conditions) plus the subset of potential decisions, which in turn depends on the subset of (direct or derived) operators of $\mathcal{H}S^2$ and the admitted values for f , \bowtie , and the initial condition K_0 .

Bagging spatial decision trees into forests

Decision trees can be generalized into sets of decision trees that operate by a majority voting policy. When

coupled with *bagging*, an ensemble learning technique that ultimately reduces the variance of the models, sets of trees are often called *random forests* and tend to be much more performing than single trees (Breiman, 2001). While they are considered to be at the verge between symbolic and functional learning, their symbolic nature is still evident: in fact, sets of trees, similar to single trees, can be analyzed and discussed, and although the process of extracting rules is not as immediate as in single trees, it is still possible (Deng, 2019; Friedman & Popescu, 2008; Meinshausen, 2010). The generalization to the forest model is relatively straightforward. While each tree is a plain decision tree, the output of a forest depends on many trees and is computed via some *voting* function. Therefore, we define a *spatial decision forest* as a pair $\mathcal{F} = (\{\tau_1, \dots, \tau_k\}, \nu)$, where $\{\tau_1, \dots, \tau_k\}$ is a collection of spatial decision trees and $\nu : \mathcal{C}^k \rightarrow \mathcal{C}$ is a *voting aggregation function*. Given a spatial forest $\mathcal{F} = (\{\tau_1, \dots, \tau_k\}, \nu)$, an image $I \in \mathcal{I}$ is classified as $\mathcal{F}(I)$ via the following definition:

$$\mathcal{F}(I) = \nu(\tau_1(I), \dots, \tau_k(I)).$$

The training algorithm for random forests differs from that of (deterministic) decision trees in many subtleties; such differences, along with the nature of the model, transform a purely symbolic method such as the decision tree into a hybrid symbolic functional approach. A first attempt toward random forests was made by Ho (1995), using the so-called *random subspace method*. Breiman's proposal (Breiman, 2001), which can be considered the standard approach to random forests, was later introduced in the *R* learning package (Liaw & Wiener, 2002), but random forests are part of a more general approach to combine several classifiers into a single one, known as *bagging*, which is still an ongoing research topic, as proven by very recent contributions (Tüysüzöğlu et al., 2022). The underlying idea behind random forests involves training all trees with no pruning condition and different trees on different subsets of the same training data. As such, in all cases, each tree is trained on a random subset of the training samples. Furthermore, at the propositional level, this also translates into using different subsets of variables for each tree. With spatial forests, one can constrain the set of available decisions by limiting the number of spatial variables, relations, feature functions and/or test operators used by each tree. The random forest learning algorithm for spatial decision tree is referred to as *Spatial RF*.

As already observed, due to the voting aggregation function, which encompasses functional computation, extracting rules from forest models is less straightforward than in the case of decision trees; however, some approaches have been proposed (Deng, 2019; Friedman & Popescu, 2008; Meinshausen, 2010). For the purpose

Table 1. Specifications for the five public datasets for land cover classification (LCC) used in the experiments.

Name	Abbr.	Sensor	# channels	Spectral coverage	Spatial resolution	Image size (px)	# labels	# classes
Indian Pines	IP	AVIRIS	200	0.4 – 2.45 μm	20 m/px	145 × 145	10249	16
Pavia University	PU	ROSIS	103	0.43 – 0.86 μm	1.3 m/px	610 × 340	42776	9
Pavia Centre	PC	ROSIS	103	0.43 – 0.86 μm	1.3 m/px	1096 × 715	148152	9
Salinas	S	AVIRIS	200	0.4 – 2.45 μm	3.7 m/px	512 × 217	54129	16
Salinas-A	S-A	AVIRIS	200	0.4 – 2.45 μm	3.7 m/px	83 × 86	5348	6

of this work, we use random forests for achieving maximal performances and decision tree models for extracting and interpreting classification rules.

Spatial C4.5 and spatial RF implementations

Implementations of decision tree/forest learning algorithms exist in several learning suites; the most popular implementations are provided by the Scikit-learn suite (Pedregosa et al., 2011) (in Python) and Weka (Witten et al., 2017) (in Java). In recent years, the Julia programming language has become increasingly popular within the scientific computing community, and although the language is still young, it offers a stable package for decision tree learning (Sadeghi, 2013). The package was extended to provide the implementations of modal versions of C4.5, among which is Spatial C4.5 and Spatial RF. The code is available as a registered Julia package hosted on a GitHub repository¹ (Pagliarini et al., 2023). The implementation makes large use of the multiple dispatch capabilities that the language enables and performs different levels of optimizations.

Spatial C4.5 is parametrized in the set of feature functions and test operators, i.e. it requires to specify the set \mathcal{FT} of feature-operator pairs and assumes that the decisions are in the following form:

$$\begin{aligned}\mathcal{S}_{prop} &= \{f(V) \bowtie v \mid (f, \bowtie) \in \mathcal{FT}, V \in \mathcal{V}, v \in \mathbb{R}\}, \\ \mathcal{S}_{mod} &= \{\langle X_1, X_2 \rangle sp \mid sp \in \mathcal{S}_{prop}\}.\end{aligned}$$

The need of including such a degree of freedom stems from the observation that some combinations feature-operator may not have a very intuitive semantics (i.e. while $\min(V) \geq v$ can be read as *V is always at least as high than v*, $\min(V) \leq v$ is interpreted as *variable V, is at least once, less than or equal to v*); in single experiments, one may want to exclude less intuitive combinations to obtain more interpretable models and, at the same time, lower the learning computation time. Now, let \mathcal{F} be the set of features that appear in \mathcal{FT} . As mentioned, the overall cost of the algorithm heavily depends on the cost of *FindBestSplitDecision*, which requires evaluating, for each image, the truth of all propositional and modal decisions on all the associated witnesses. We optimized the memory accesses performed in this step, by means of additional data structures used to maximize the locality of the

memory accesses, as well as to reduce recomputations. Briefly, we obtain such as result as follows. The input dataset, viewed as a structure $\mathbb{R}^{W \times H \times |\mathcal{V}| \times |\mathcal{I}|}$, is first processed so that all features for all rectangles are computed. This information is then stored in a hash table $\mathcal{H}_{prop} : \mathbb{K}(\mathbb{D}^2) \times \mathcal{I} \times \mathcal{F} \times \mathcal{V} \rightarrow \mathbb{R}$, which essentially encodes the valuation function, and can be used for checking the truth of both propositional and modal decisions:

$$\begin{aligned}I, r \models f(V) \bowtie v &\Leftrightarrow \mathcal{H}_{prop}[r, I, f, V] \bowtie v, \\ I, r \models \langle X_1, X_2 \rangle f(V) \bowtie v &\Leftrightarrow \exists s \text{ s.t. } (r, s) \in R_{X_1, X_2}, \\ &\quad \mathcal{H}_{prop}[s, I, f, V] \bowtie v.\end{aligned}$$

Observe how, while checking a propositional decision only requires a single lookup operation plus the evaluation of a condition, in the worst case, checking a modal decision using \mathcal{H}_{prop} requires performing a propositional check for a number of rectangles that is $\mathcal{O}(|\mathbb{K}(\mathbb{D}^2)|)$; however, we can alleviate this cost by exploiting structural patterns. First, leveraging the linear order onto which \bowtie is defined, the last expression can be written as follows:

$$I, r \models \langle X_1, X_2 \rangle f(V) \bowtie v \Leftrightarrow \text{aggr}_{(r,s) \in R_{X_1, X_2}}(\mathcal{H}_{prop}[s, I, f, V]) \bowtie v,$$

where the function *aggr* (i.e. *aggregate*) is max when $\bowtie \in \{>, \geq\}$ and min when $\bowtie \in \{<, \leq\}$. The result of the aggregation consists of a real number, and it can, therefore, be reused for checking, on the same image I and rectangle r , the truth of any other modal decision with the same X_1, X_2, f, V, \bowtie , but different value for v . Following such an intuition, we introduce a second hash table $\mathcal{H}_{mod} : \mathbb{K}(\mathbb{D}^2) \times \mathcal{I} \times \mathcal{FT} \times \mathcal{V} \times \mathcal{X}^2 \rightarrow \mathbb{R}$ to carry the results of all aggregations and therefore allow a more efficient truth checking of modal decisions:

$$I, r \models f(V) \bowtie v \Leftrightarrow \mathcal{H}_{mod}[r, I, (f, \bowtie), V, X_1, X_2] \bowtie v.$$

Clearly, for a fixed function f , all else being equal, \mathcal{H}_{mod} may store different thresholds for $(f, >)$ and $(f, <)$, while \mathcal{H}_{prop} would store a unique threshold for any value of \bowtie ; moreover, \mathcal{H}_{mod} is much larger in size than \mathcal{H}_{prop} . Finally, the same design pattern can be adapted to the case of global decisions by introducing a different hash table $\mathcal{H}_{glob} : \mathcal{I} \times \mathcal{FT} \times \mathcal{V} \rightarrow \mathbb{R}$, which can be constructed using similar aggregations:

¹<https://github.com/aclai-lab/ModalDecisionTrees.jl>.

$$I, r \Vdash \langle G \rangle f(V) \bowtie v \Leftrightarrow \mathcal{H}_{glob}[I, (f, \bowtie), V] \bowtie v.$$

Similar to \mathcal{H}_{mod} , the values in \mathcal{H}_{glob} depend on \bowtie but, because the truth of global decisions does not depend on r , this structure is smaller and more scalable than the previous ones (observe that in typical, real-world scenarios the size of the image has a much heavier impact than the number of test operators). Exploratory analysis shows that the best performances are attained by precomputing \mathcal{H}_{prop} and \mathcal{H}_{glob} prior to the learning phase and by performing memoization on \mathcal{H}_{mod} throughout the learning process.

Observe that, to avoid recomputations, both the precomputed and memoized values can be shared across the training of different tree models. This is especially useful when training forests of hundreds/thousands of trees, and can ease the computational load needed to find the optimal hyperparametrization for the learning algorithms.

Experimental evaluation

In this section, we consider the task of LCC and evaluate the capabilities of spatial decision trees. We report and discuss the results of our experiments, analyzing them under three different directives, namely, their statistical performance, their complexity and their interpretability.

Data

We considered five datasets that are commonly used to benchmark neural network-based methods for LCC (see Ahmad et al., 2017; Audebert et al., 2019; Cao et al., 2018; Hong et al., 2022; Hu et al., 2015; Jiang et al., 2019; Lee & Kwon, 2017; Li et al., 2017; Makantasis et al., 2015; Mou et al., 2017; Roy et al., 2020; Santara et al., 2017). The datasets are known as *Indian Pines*, *Pavia University*, *Pavia Centre*, *Salinas* and *Salinas-A*, respectively; a summary of their characteristics is reported in Table 1. Each dataset consists of a hyperspectral image, or *scene*, of a piece of land coupled with a *ground-truth label mask*, providing the correct class for some of the pixels in the scene. In all cases, the scene is captured by a dedicated sensor during a flight campaign; specifically, Pavia University and Pavia Centre were collected using a ROSIS sensor (Reflective Optics System Imaging Spectrometer) and the remaining ones using an AVIRIS sensor (Airborne Visible/Infrared Imaging Spectrometer). The ROSIS and AVIRIS yield spectral detections covering a range of frequencies from $0.43\mu\text{m}$ to $0.86\mu\text{m}$ and from $0.4\mu\text{m}$ to $2.45\mu\text{m}$, with a number of channels of 103 and 200, respectively. Note that, in the case of hyperspectral images, there exists an implicit order

between the variables, and in fact, it is often the case that hyperspectral channels with close wavelengths are correlated. The size of the scenes varies from 86×83 pixels for Salinas-A to 1096×1096 pixels for Pavia Centre; note, however, that not all pixels are labeled.

The typical approach to LCC involves collecting either all labeled pixels or a randomly sampled subset and applying a learning algorithm for binary or multi-class classification. In some cases, authors have leveraged the spatial structure in a limited way, by considering, for the classification of each single pixel, a set of neighboring pixels; the neighborhood is generally in the form of a $d \times d$ window centered in the pixel, for a natural odd number d . We adopted the same solution and extracted, for each dataset, a collection of $d \times d$ labeled images (one for each labelled pixel). All five datasets originally presented a class imbalance, with some classes appearing thousands of times, while others only appearing a few dozen times. For the case of entropy-based decision tree learning, such imbalances generally cause biases toward the most occurring classes, and since our objective is to extract logical descriptions that are equally accurate for all classes, we chose to level out this imbalance. Thus, a fixed number P of pixels were randomly sampled for each class, using upsampling for classes with less than P samples. The experiments were conducted in a randomized cross-validation setting, in which this sampling step is performed n times, and each time the set is partitioned into two balanced sets, one for training the model and the other for testing and evaluating it. Within this context, we set $d = 3$, $P = 100$ and $n = 10$ to produce, for each dataset, 10 balanced subdatasets of 3×3 images. Furthermore, a 80%–20% balanced split was performed when splitting each subdataset into training and test, using a policy for ensuring that they were *strongly disjointed*. Such a policy prevents any two images with non-empty overlap on the original scene to end up on different sides of the training-test frontier, in order to avoid *data leakage*, which occurs when part of the information in the test set appears in the training set as well, biasing the algorithm and, ultimately, affecting the performance estimation.

A quick overview of the literature displays a wide variety of experimental practices and evaluation settings when dealing with these datasets; however, only a few works seem to prevent data leakage (Audebert et al., 2019) and perform a balancing pre-process step (Cao et al., 2018), which may or may not affect the evaluation of the learned models. Having said that, in all cases, neural models achieved nearly optimal performances: for Indian Pines and Pavia University, the average accuracy is in the range ~ 79 – 99% (Audebert et al., 2019; Cao et al., 2018; Hu et al., 2015; Lee & Kwon, 2017; Li et al., 2017; Makantasis et al., 2015;

Mou et al., 2017; Roy et al., 2020; Santara et al., 2017); for Pavia Centre, the average accuracy is at $\sim 99\%$ (Makantasis et al., 2015); for Salinas, the average accuracy is in the range of $\sim 92\text{--}99\%$ (Hu et al., 2015; Jiang et al., 2019; Lee & Kwon, 2017; Makantasis et al., 2015; Roy et al., 2020; Santara et al., 2017) and for Salinas-A, the average accuracy is at $\sim 97\%$ (Ahmad et al., 2017). Also, propositional random forest models attained average accuracies in the range of $\sim 69\text{--}71\%$ (Mou et al., 2017) on Indian Pines and Pavia University, and around $\sim 93\%$ on Salinas (Jiang et al., 2019).

Experimental setting

From each dataset, we extracted symbolic models using traditional C4.5 and its random forest generalization, as well as Spatial C4.5 and Spatial RF under different parametrizations; then, we compared both their performances and the models themselves, with the aim of showing how taking into account the spatial component improves the result of a traditional symbolic approach in the context of the LCC task. For each training-test split, different approaches to decision tree modeling are deployed and compared. We separate the approaches to symbolic learning for LCC into *pure* approaches, where the models are learned on the 3×3 image itself, and *derived* ones, where the input image is processed beforehand using functional filters. The pure approaches included in this experiment are referred to as *single-pixel* propositional approach, in which standard C4.5 is applied to the numerical description of the pixel to be classified (i.e. the central pixel), disregarding the neighboring pixels, *flattened* propositional approach, in which standard C4.5 is applied to the numerical descriptions of all pixels in the 3×3 image, disregarding the spatial structure, and *RCC8* (resp., *RCC5*) modal approaches, in which Spatial C4.5 with \mathcal{HS}_{RCC8}^2 (resp., \mathcal{HS}_{RCC5}^2) is applied to the 3×3 image, and the image channels are regarded as spatial variables. Each of the above settings has a corresponding derived one, obtained by applying, before training, a 3×3 average convolutional filter. In derived settings, the outer 5×5 box around each 3×3 image is first used to compute a 3×3 *average image* which is, then, fed *as is* to the learning algorithm of the corresponding pure setting; in this way, each pair of approaches is immediately comparable. To fix the ideas, consider the case of *Indian Pines*, which has 200 channels. The pure single-pixel approach trains trees that make decisions on the 200 featured values of the central pixel; on the other hand, in the corresponding derived setting (referred to as *avg*), trees are trained from descriptions of 200 values that are, each, the average of a channel within the

3×3 neighborhood. Instead, trees trained within the pure flattened approach make decisions on the $3^2 \times 200 = 1800$ featured values within the 3×3 image, while the corresponding derived setting (*avg+flattened*) uses $3^2 \times 200 = 1800$ values resulting from the 3×3 average convolution performed on the 5×5 outer box. As for the spatial approaches, they are always applied on 3×3 images, except in the derived cases (*avg+RCC8* and *avg+RCC5*) the image is the result of the convolution. It is of note that, because LCC is invariant under rotation and reflection, spatial approaches are more likely to grasp complex patterns when using topological relations, as opposed to directional ones.

All experiments were run using the Julia package ModalDecisionTrees.jl and the Sole.jl framework.² From an operational perspective, each approach differs from the others by how data are preprocessed prior to the learning phase and by the algorithm parametrization. As a matter of fact, when applied to datasets of 1×1 images, Spatial C4.5 behaves exactly as traditional C4.5, thus the same implementation of Spatial C4.5 presented in the previous section can be used for implementing traditional and spatial approaches. With regard to pre-pruning, different parametrizations were tested using the single-pixel baseline approach, and the one achieving the highest cross-validation performance was finally fixed. The fixed pre-pruning conditions are minimum number of samples per leaf of 4, minimum information gain of 0.01 for a split to be meaningful and maximum entropy at each leaf of 0.3. As for the parametrization of the spatial approaches, we fixed r_0 to be the minimum rectangle that contains the central pixel (i.e. the pixel to be classified). Moreover, the set of decisions was induced by setting:

$$\mathcal{FT} = \{(max, \leq), (max, \geq), (min, \leq), (min, \geq)\},$$

and global decisions were disallowed throughout the learning. A similar grid search for the parametrization of the random forest models was performed; ultimately, the number of trees was fixed to 100, with each tree being trained on a random subset of samples with a size of 70% relative to the whole set, and only 50% of the total variables.

Ultimately, considering that each approach is tested in both the single-tree and the random forest versions, this setting involves comparing $5 \times (4 \times 2) \times 2 = 80$ different approaches. As mentioned above, cross-validation with 10 repetitions was deployed; therefore, a total of 800 models were trained. Each model was evaluated via standard performance metrics for multi-class classification, namely, *overall accuracy*, κ *coefficient* (which relativizes the accuracy to the probability

²<https://github.com/acalai-lab/Sole.jl>.

of a random answer being correct Cohen, 1960), $\text{cohen1960coefficient}$) and *mean precision*. With balanced test sets, the overall accuracy also corresponds to the *mean recall*.

Statistical performance

Overall, the total computation time needed for the experiments was around 523 h (about 21 days) on a dual-socket Dell PowerEdge server with Intel Xeon Gold 6238 R processors, using multi-threading with a number of threads varying from 26 to 32. Table 2 shows the statistical cross-validation results of pure and derived approaches applied to the five datasets; for each approach and dataset, the average and standard deviation of κ coefficient, overall accuracy and mean precision are reported. The discussion that follows is mainly based on the κ coefficient, but the other two metrics reveal similar insights. It can, first, be observed that the pure and derived flattened approaches lead to a severe performance degradation; the average κ coefficient hovers around 0% in all cases, and it is below 0% in most cases. This is likely due to the large number of decisions, which has a negative effect on the greedy learning strategy; as such, these two approaches can be excluded from a relevant method comparison. Pure approaches tend to attain lower performances than their derived counterpart; this is always the case for the single-pixel approach, while it holds for the spatial approaches in 14 out of 20 cases. Across all datasets, the best performances in the pure and derived cases are always attained with spatial approaches, which, therefore, appear to yield consistently better results than propositional ones. The improvements seem to be proportional to the intrinsic hardness of the classification dataset; compared with the single-pixel approach, the average improvement of spatial approaches ranges from about 1% point (Salinas-A, decision tree) to 8.5% points (Indian Pines, random forest). From a qualitative perspective, there does not seem to be a clear winner between the two topological logics: RCC8 and RCC5 behave quite similarly, with the greatest difference in terms of κ being as low as 2% points (Indian Pines, derived spatial approaches). However, across the four tree types, RCC8 yields the best results in 8 out of 20 cases, while RCC5 in 14 out of 20 cases. As expected, random forest models always achieve higher performances when compared with their single-tree counterpart, with the only two exceptions represented by the derived spatial approach; apart from these isolated cases, which are likely due to fluctuations, the improvement ranges from 12% points (Indian Pines, pure single-pixel) to 1% point (Salinas-A, derived single-pixel). Figure 5a graphically shows the distribution of the average accuracies for the random forest models. As it appears, the spatial approaches tend to

achieve accuracies of $\sim 90\text{--}96\%$ for Indian Pines and Pavia Centre, $\sim 85\text{--}95\%$ for Pavia University, $\sim 93\text{--}98\%$ for Salinas and $\sim 96\text{--}100\%$ for Salinas-A. Although a proper comparison with the existing literature on the same datasets is hampered by the differences in the experimental setting, we point out that these results are consistent with those from the neural literature and are also consistently better than those attained by other works using random forest models (refer to the *Data* section above).

In order to assess whether the performances of spatial approaches are significantly better than propositional ones, the validation accuracies throughout the 10 repetitions can be interpreted as statistical populations and subjected to statistical hypothesis tests. Each accuracy population, identified by dataset and approach, is represented via a boxplot in Figure 5 (top), which provides an immediate graphical overview of the performances. In this context, we rely on the Holm-adjusted Wilcoxon signed-rank test, as explained in by Benavoli et al. (2016), and report the results in the form of *Critical Difference diagrams* (Demšar, 2006), which is a standard methodology for such comparisons. Figure 5 (bottom) shows two sets of Critical Difference diagrams comparing approaches based on pure random forests and derived random forests, respectively. Each diagram refers to a single dataset and shows the average rank attained with each approach throughout the 10 repetitions (smaller is better); underneath each diagram, bars are shown, which connect and group approaches that are statistically indistinguishable. The two sets of diagrams reveal virtually the same trends, and a few observations can be made. In all cases, spatial approaches have better ranks than single-pixel ones and are always in the group of significance with the lowest rank. Furthermore, for Indian Pines and Pavia University, both spatial approaches show a significant performance improvement over the single-pixel approach, both in the pure and the derived case; across all datasets, the improvement is considered significant in 7 out of 10 cases.

Model complexity

With the aim of analyzing the structural and computational complexities of the models and confronting them with their statistical performance, we report in Table 3 some additional metrics for each case, including the number of rules per decision tree trained and the average training time of the model. The five datasets can be ordered by complexity by considering the number of rules required for classifying each one of them. Ignoring the flattened approaches, which are

Table 2. Cross-validation results on five datasets using different approaches based on propositional and spatial decision trees (in bold). For each approach, the average and the standard deviation across 10 repetitions of the κ coefficient, the overall accuracy and the mean precision are reported in percentage points. For each result line, the average performance of the best pure approach and the best derived approach is highlighted.

Dataset	Pure approaches										Derived approaches																													
	Single-pixel					flattened					RCC8					RCC5					avg					avg+flattened					avg+RCC8					avg+RCC5				
	κ	acc	prec	prec	κ	acc	prec	prec	κ	acc	prec	κ	acc	prec	κ	acc	prec	κ	acc	prec	κ	acc	prec	κ	acc	prec	κ	acc	prec	κ	acc	prec								
Random Forest	IP	avg	82.9	83.9	84.6	0.5	6.7	6.4	91.5	92.0	92.6	91.2	91.8	92.4	86.8	87.6	88.2	-1.9	4.4	4.7	92.0	92.5	93.1	94.0	94.4	94.8														
		std	2.2	2.0	1.9	1.5	1.4	1.0	2.3	2.2	2.2	2.2	1.9	1.8	1.7	3.1	2.9	2.9	2.0	1.9	3.6	2.2	2.1	2.0	2.8	2.7	2.5													
	PU	avg	81.2	83.3	84.4	-3.5	8.0	11.4	87.2	88.7	89.8	87.9	89.2	90.0	83.4	85.2	86.6	-4.5	7.1	9.9	87.8	89.1	90.0	87.9	89.2	90.3														
		std	4.0	3.5	4.1	5.3	4.7	6.7	4.1	3.7	3.0	5.5	4.9	4.4	2.8	2.5	2.3	4.3	3.8	5.9	3.7	3.3	3.3	3.9	3.5	3.4														
	PC	avg	88.1	89.4	90.6	-5.5	6.2	12.8	92.0	92.9	93.5	92.2	93.1	93.6	91.8	92.7	93.3	-6.8	5.1	9.0	92.6	93.4	94.0	92.6	93.4	94.0														
		std	4.5	4.0	3.5	4.4	3.9	7.1	3.2	2.9	2.7	3.2	2.9	2.6	3.2	2.8	2.6	5.1	4.5	10.5	2.8	2.5	2.3	3.1	2.7	2.6														
Decision Tree	S	avg	92.9	93.3	93.6	-4.1	2.4	3.2	94.7	95.1	95.4	94.1	94.5	94.8	94.5	94.9	95.3	-3.9	2.6	4.8	95.8	96.1	96.3	95.9	96.2	96.4														
		std	1.4	1.3	1.3	1.5	1.4	2.8	1.7	1.6	1.5	1.4	1.3	1.4	1.3	1.2	1.1	1.7	1.6	2.1	1.3	1.2	1.2	1.5	1.4	1.4														
	S-A	avg	97.0	97.5	97.8	-4.2	13.2	11.3	98.8	99.0	99.1	98.6	98.8	98.9	97.6	98.0	98.2	-5.4	12.2	11.8	98.4	98.7	98.8	98.0	98.3	98.5														
		std	2.7	2.3	2.0	8.5	7.1	5.9	1.7	1.4	1.3	2.1	1.8	1.6	2.8	2.3	2.2	8.7	7.2	6.1	2.8	2.3	2.1	3.0	2.5	2.3														
	IP	avg	70.9	72.7	72.8	0.2	6.4	6.5	78.3	79.6	80.5	79.4	80.7	81.6	76.7	78.2	79.5	0.6	6.8	7.6	80.5	81.7	82.9	81.3	82.5	83.6														
		std	4.1	3.8	4.3	2.5	2.3	1.8	3.4	3.1	3.1	3.2	3.0	3.1	3.7	3.4	3.3	1.8	1.7	2.0	4.1	3.9	4.0	2.7	2.6	3.4														
Random Forest	PU	avg	71.6	74.8	75.9	-0.9	10.3	10.8	77.5	80.0	80.6	77.6	80.1	80.6	73.9	76.8	78.2	-1.8	9.6	10.2	76.9	79.4	80.5	77.5	80.0	81.0														
		std	4.3	3.9	3.6	5.7	5.0	4.8	7.5	6.7	7.3	6.6	5.8	6.6	5.5	4.9	5.5	5.8	5.2	6.1	4.6	4.1	4.8	4.6	4.1	4.6														
	PC	avg	84.9	86.6	88.3	-1.8	9.6	9.8	89.1	90.3	91.5	89.4	90.6	91.6	89.8	90.9	91.7	-2.6	8.8	11.4	88.9	90.1	91.0	88.8	90.0	90.9														
		std	4.1	3.7	2.6	5.8	5.2	4.1	3.1	2.8	2.5	2.9	2.6	2.3	3.5	3.1	2.6	5.4	4.8	7.5	4.2	3.8	3.3	4.1	3.6	3.2														
	S	avg	88.5	89.2	89.9	0.2	6.4	6.2	91.2	91.8	92.2	90.8	91.4	91.9	92.5	93.0	93.6	-1.4	4.9	4.4	93.2	93.6	94.1	93.3	93.8	94.3														
		std	2.4	2.3	2.2	1.5	1.4	1.7	1.9	1.8	1.7	2.2	2.1	2.0	1.4	1.3	1.4	2.4	2.3	2.6	2.6	2.4	2.2	2.2	2.1	1.9														
Decision Tree	S-A	avg	95.4	96.2	96.6	-3.6	13.7	12.4	96.4	97.0	97.4	96.4	97.0	97.4	96.6	97.2	97.5	-5.2	12.3	13.0	98.6	98.8	98.9	98.6	98.8	98.9														
		std	3.4	2.8	2.5	9.7	8.1	7.1	3.0	2.5	2.2	3.0	2.5	2.2	3.0	2.5	2.2	7.2	6.0	9.0	2.7	2.2	2.0	2.7	2.2	2.0														

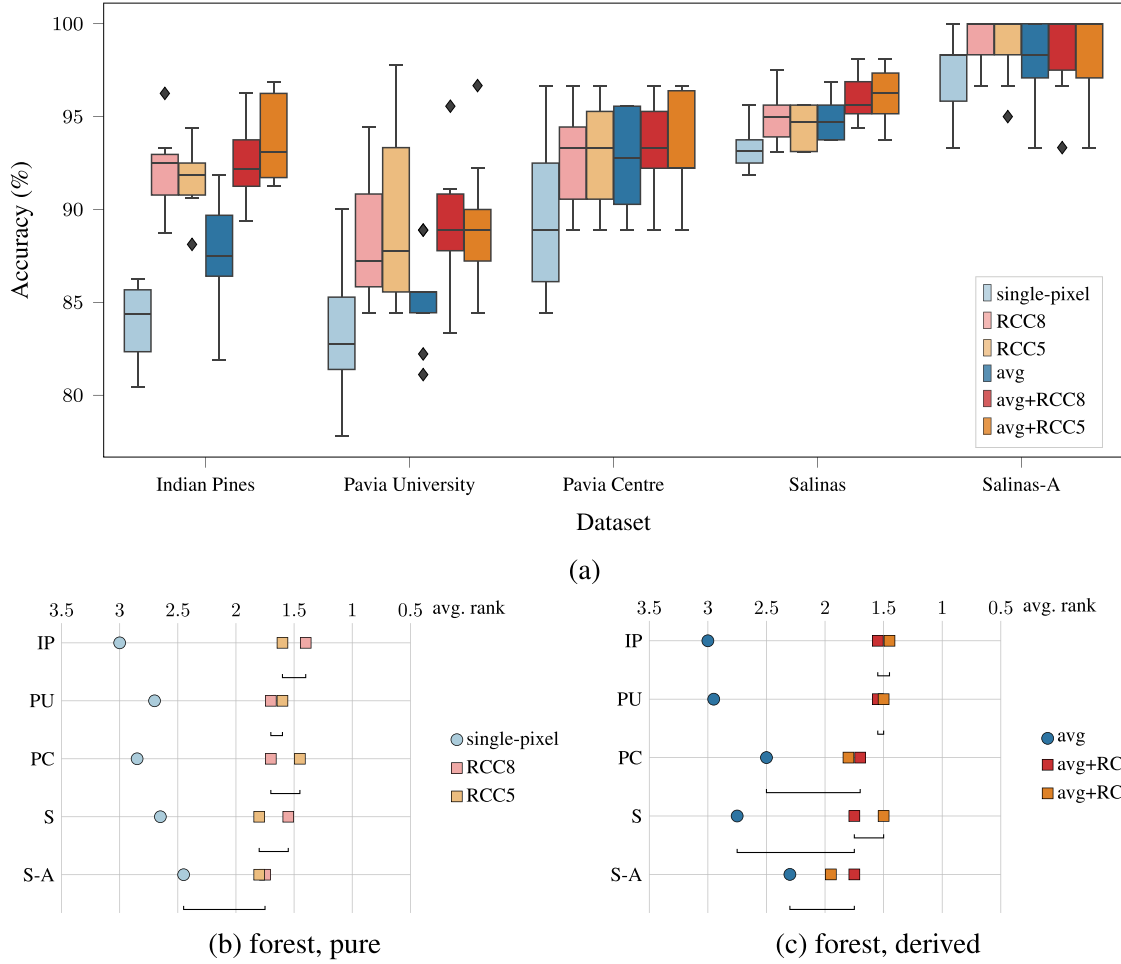


Figure 5. Hypothesis test results on the average accuracies of the eight approaches and five datasets. At the top (a), the distributions of the average accuracies for the random forest models are shown via a boxplot, reporting a five-number summary (minimum, first quartile, median, third quartile and maximum) and the outliers (i.e. observations farther away than 1.5 the interquartile range). At the bottom, critical difference diagrams comparing the pure (b) and derived random forest (c) approaches.

not comparable in terms of performance, Indian Pines encloses the task that causes the trained trees to be the largest in 11 out of 12 cases (i.e. except for forest avg); in 9 out of the remaining 11 cases (i.e. except for single-tree RCC8 and forest avg+RCC8), Pavia University, Salinas, Pavia Centre and Salinas-A follow, in this order. With the exception of derived single-tree approaches on Salinas and derived random forests on Indian Pines, spatial trees are, on average, smaller than their (non-flattened) propositional counterparts. As for the two exceptions, the single-tree avg approach on Salinas yields an average number of leaves equal to 19.4, which is not too smaller than that of avg+RCC8 (22.7) and avg+RCC5 (23.3), while forest avg on Indian Pines yields substantially smaller trees (50.6 leaves) compared with the spatial counterparts (73.3 and 78.2 leaves) but also has a lower κ (86.8%, compared with 92.0% and 94.0%); a possible cause could be the pruning condition causing the algorithm to stop the learning prematurely. When comparing pure and derived approaches, the avg approach always yields smaller trees than the propositional approach, with a gap ranging from 0.4 to 10.9 (single tree on S-A and single tree on PU, respectively); as for the spatial

approaches, the same pattern only holds 9 out of 12 cases. When comparing random forests and decision trees, only in 4 out of 40 cases (single-pixel, RCC8, RCC5, and avg, all on Indian Pines) the average size of the trees in forest models is less than that of the respective single decision tree model; this is expected, given that forest models do not involve pruning conditions. Fixed a dataset, in general, with larger values for κ , the number of rules seems to decrease; this rule of thumb does not hold in all cases in the table, but in those cases where it does not, the differences are of a few decimals (e.g. pure RCC5 vs. pure RCC8 forests on Salinas). A similar trend also emerges in the fact that, with only 3 exceptions out of 20 (derived forest on Indian Pines, derived single tree on Pavia Centre and derived single tree on Salinas), for any propositional (pure or derived) approach, there always exists a corresponding spatial approach (pure or derived) that yields trees that are both smaller and more accurate.

With respect to the training times, it can be observed how the spatial approaches require much more time than propositional ones. With respect to propositional approaches, on average across all datasets, learning a tree with RCC8 requires 31 (derived

case) to 46 (pure case) times more time, and RCC5 requires 20 (derived case) to 25 (pure case) times more time. Considering that the performances and structural complexities of RCC8 and RCC5 trees are similar, it can be argued that RCC5 provides an optimal trade-off between expressiveness and computational complexity. It can also be observed that training forests of 100 trees always takes less time than a hundred times the time of training a corresponding single tree. This could be due to the (previously observed) fact that, despite forests having no pruning condition, in most cases, bagged trees are on average smaller than single trees trained on the same data. However, it is likely that the shared memoization policy has a much stronger impact, with regard to this point.

Post-hoc interpretation

As mentioned above, the statistical performances of a symbolic method is a limited metric for evaluating its usefulness; given the interpretable nature of decision trees, on top of the above analysis, trained models can be inspected to gain insights into the extracted knowledge. A natural way to carry out this kind of analysis is to choose a performant model, analyze its structure and its classification rules and relate their accuracy to the performance of the model. Within our experimental context, we choose a propositional baseline tree and a spatial tree trained on the same dataset, opting for pure approaches, which are more easily interpretable.

We begin this exploration by analyzing, for the propositional and the spatial cases, a pair of single trees trained on the same training-test split of Salinas-A, where the classification task is to distinguish six different types of vegetables. Figure 6 (top) shows two trees extracted with the single pixel, and with the RCC8 approach, respectively. As for their statistical performances, the single-pixel tree displays a validation κ of 92%; however, we found that the majority of the misclassifications, in this case, occur between two specific classes: 25% of the test samples

belonging to class C_2 (*corn senesced green weeds*) are misclassified by the model as belonging to class C_6 (*lettuce romaine, 7 weeks*). The spatial tree, on the other hand, has 100% accuracy (and κ), and, with respect to the single-pixel model, it increases the recall of class C_2 from 75% to 100%, while also reaching optimal performance for the other five classes. These trees have the same topology but only differ by the split decisions. The structural similarity allows for an additional comparison of the two trees, which reveals that 8 out of 10 split decisions involve variables with indices that are within three units apart (e.g. in the two cases, the decisions at the root node depend on V_{42} and V_{43} , respectively). Considering that hyperspectral channels with similar wavelengths are correlated, and given that Salinas-A has 200 variables, this suggests that the natural patterns they are capturing for each class are quite similar. Also, both trees have exactly one leaf (and, thus, one classification rule) for each of the six classes C_i ; therefore, for each class, they encompass a single logical condition $\varphi_{C_i} = \varphi_{\ell_i}$ that is (in a statistical sense) both necessary and sufficient for C_i . The two conditions that emerge for class C_2 are shown in Figure 6 (bottom) in their simplified form. At a glance, the formulas have overlapping semantics, different structures and use slightly different variables. The single-pixel, propositional condition plainly translates to *within the pixel, V_{42} is less than 3367, V_{85} is at least 1296, and V_7 is at most 1628*. The spatial one (in this case, a \mathcal{HS}_{RCC8}^2 -formula) describes a richer spatial scenario, which can be translated in natural language as follows: *within 3×3 window surrounding the pixel, the minimum of V_{43} is less than 1978, and there exists at least one rectangle of neighboring pixels in which all of the pixels have $V_{83} \geq 638$ and $V_6 \leq 1625$* . In this case, the misclassifications of C_2 samples by the single-pixel tree can be attributed (at least) to the first single-pixel split; this split is performed on the characteristics of the pixel itself, while the class C_2 , seemingly, requires spatial considerations on the neighboring pixels to be correctly distinguished from C_6 .

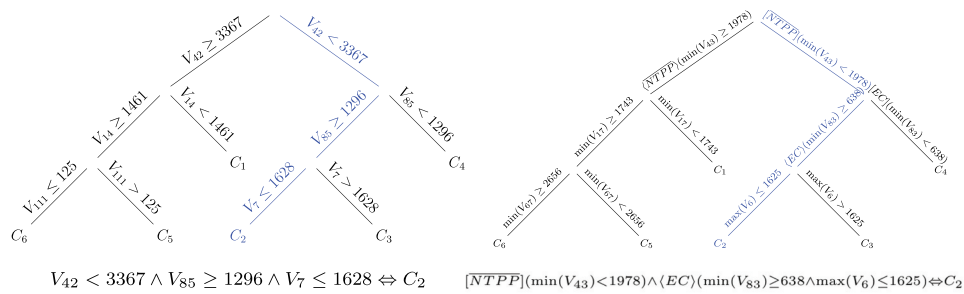


Figure 6. Comparison between a propositional (left) and a spatial (right) decision tree extracted on Salinas-A. The highlighted paths show the rules that each model has extracted for the class C_2 (i.e. *corn senesced green weeds*).

Table 3. Cross-validation results on five datasets using different approaches based on propositional and spatial decision trees (highlighted in bold). For each approach, the average across 10 repetitions of the following metrics is reported: κ coefficient (percentage points), number of leaves/rules per tree r_t and training time t (measured in seconds and divided by the number of threads). For both the random forest and the decision tree, the average of each performance measure across the five datasets is also shown (avg), as well as the overall average values summarizing the performance of both models (ov. avg). For each result line, the average performance of the pure approach and the derived approach with the smallest number of rules is highlighted.

Dataset	Pure approaches															Derived approaches																							
	single-pixel					flattened					RCC8					RCC5					avg					avg+flattened					avg+RCC8					avg+RCC5			
	κ	r_t	t	κ	r_t	t	κ	r_t	t	κ	r_t	t	κ	r_t	t	κ	r_t	t	κ	r_t	t	κ	r_t	t	κ	r_t	t	κ	r_t	t	κ	r_t	t						
Random Forest	IP	82.9	56.6	39	0.5	153.5	353	91.5	40.2	87.4	91.2	40.4	351	86.8	50.6	16	-1.9	153.3	140	92.0	73.3	869	94.0	78.2	580														
	PU	81.2	54.4	7	-3.5	149.1	59	87.2	37.6	215	87.9	38.1	144	83.4	52.9	7	-4.5	159.3	64	87.8	40.2	207	87.9	40.6	128														
	PC	88.1	35.9	10	-5.5	147.5	124	92.0	23.8	312	92.2	23.8	188	91.8	29.4	9	-6.8	155.7	111	92.6	24.8	188	92.6	25.1	114														
	S	92.9	46.4	28	-4.1	334.8	385	94.7	33.7	2547	94.1	33.6	1342	94.5	38.9	29	-3.9	338.5	333	95.8	31.5	855	95.9	31.4	507														
	S-A	97.0	9.0	3	-4.2	82.8	46	98.8	7.5	128	98.6	7.5	79	97.6	8.6	4	-5.4	81.7	59	98.4	6.6	145	98.0	6.5	90														
avg	88.4	40.5	17	-3.4	173.5	193	92.8	28.6	815	92.8	28.7	421	90.8	36.1	13	-4.5	177.7	141	93.3	35.3	453	93.7	36.4	284															
Decision Tree	IP	70.9	66.0	3	0.2	255.0	34	78.3	50.1	123	79.4	51.0	82	76.7	56.9	4	0.6	247.7	47	80.5	46.1	99	81.3	47.0	99														
	PU	71.6	23.1	1	-0.9	109.2	10	77.5	22.0	24	77.6	22.1	19	73.9	34.0	1	-1.8	117.3	12	76.9	23.8	19	77.5	24.6	15														
	PC	84.9	18.0	2	-1.8	117.4	26	89.1	13.5	46	89.4	13.5	29	89.8	15.4	2	-2.6	129.5	21	88.9	13.8	15	88.8	13.7	11														
	S	88.5	22.1	4	0.2	264.1	47	91.2	23.2	296	90.8	21.7	241	92.5	19.4	5	-1.4	248.4	66	93.2	22.7	70	93.3	23.3	59														
	S-A	95.4	7.9	0	-3.6	85.5	7	96.4	6.8	22	96.4	6.8	20	96.6	6.9	1	-5.2	78.7	13	98.6	6.0	22	98.6	6.0	19														
avg	82.3	27.4	2	-1.2	166.2	25	86.5	23.1	102	86.7	23.0	78	85.9	26.5	3	-2.1	164.3	32	87.6	22.5	45	87.9	22.9	41															
ov. avg	85.3	33.9	10	-2.3	169.9	109	89.7	25.8	459	89.8	25.8	250	88.4	31.3	8	-3.3	171.0	87	90.5	28.9	249	90.8	29.6	162															

With the aim of both evaluating the generalization ability of these methods and further deepening their practical usefulness, as a last set of experiments, we perform a variation of this study, but considering, this time, the full extent of one of the datasets. We carry out this study on Indian Pines, which is the hardest of the five datasets, and, since for this dataset, RCC5 is, on average, more performant than RCC8, we use this approach for obtaining the spatial tree. A single-pixel tree and an RCC5 spatial tree are trained on a 20% slice of the whole dataset (10249 labelled samples), which encompass 1993 samples, with a class imbalance that approximates the original class distribution. Since the RCC5 model requires a 3×3 surrounding neighborhood, we ignore the labelled pixels that are on the edge of the scene, for which the neighborhood of pixels is not available; this leads to discarding 255 of the 10,249 original labelled pixels, leading to a subdataset of 9994 samples. Starting, again, from the statistical performances, when tested on the full set of samples, the two models achieved κ values equal to 50% and 58%, an overall accuracy of 58.0% and 64.8%, and an average precision of 64.2% and 72.4%, respectively. The relative improvement achieved by the RCC5 approach with respect to the single-pixel approach is still evident. However, with respect to the performances of the previous set of experiments (Table 2), where the κ for the same models was 70.9% and 78.3%, respectively, a performance degradation affects both methods; this is probably due to the fact that the dataset is almost eight times larger (9994 samples, compared with the previous training size of 1280 samples), and the fraction of samples used for training is much smaller (20%, compared with 80%). Moreover, the training and test sets are now both unbalanced, which ultimately makes these results uncomparable with the previous ones. Table 4 shows a detailed report of the performance of the two trees in terms of confusion matrices, per-class recalls and precisions. Across the whole dataset, the single-pixel tree scores a 0% recall on five classes (C_1 , C_7 , C_9 , C_{10} , C_{16}); the spatial tree, for one of these classes (C_9), achieves a 35% recall by correctly classifying 7 out of 20 instances, while it does not improve the recalls of the other four classes. As for the remaining classes, in 10 out of 12 classes, the spatial tree achieves higher recall, with improvements that are sometimes as large as 34.7 (C_{15}) and 29.3 (C_5) percentage points; in the remaining two cases, the gap is smaller (less than 1.9 percentage points). Similar observations can be made by inspecting the per-class precisions.

Table 4 also reports the total and per-class number of rules. The total number of leaves/rules for the

single-pixel and the RCC5 tree is 28 and 25, respectively, which is, again, in line with our considerations about the relation between statistical performance and structural complexity. For the two trees, it is the case that for those classes for which the tree scores a 0% recall (C_1 , C_7 , C_{10} , C_{16} for both trees, and C_9 for the single-pixel case), the tree has exactly zero classification rules (and, therefore, an undefined precision). Additionally, for three classes (C_3 , C_6 , C_{11}), the single-pixel tree provides a smaller (non-zero) number of associated rules, and for six classes (C_2 , C_4 , C_5 , C_9 , C_{14} , C_{15}), the spatial tree does. Focusing on C_{15} , the single-pixel tree and the RCC5 tree have two rules and one rule for this class, respectively. By inspecting the trees, we find that the two propositional, single-pixel rules for C_{15} , whose corresponding paths are shown in Figure 7, have a support, confidence and conviction of 0.3%, 53.6%, 1705% and 208.6%, and 0.6%, 48.4%, 1540% and 187.7%, respectively. The two rules are:

$$\begin{array}{ll} V_{189} \leq 844 \wedge & V_{189} \leq 844 \wedge \\ V_{27} \leq 12044 \wedge & V_{27} > 12044 \wedge \\ V_{28} > 10592 \wedge & V_{101} \leq 6972 \wedge \\ V_{36} \leq 25966 \wedge & V_{35} \leq 21123 \wedge \\ V_{38} \leq 38474 \Rightarrow C_{15}, & V_{102} > 3460 \wedge \\ & V_{39} \leq 33722 \Rightarrow C_{15}. \end{array}$$

On the other hand, the spatial tree has a single rule for the same class with a support of 2%, a confidence of 77.0%, a lift of 2451% and a conviction of 421.1%. The rule is:

$$\begin{array}{l} [DR] \min(V_{178}) < 1749 \wedge \\ [DR] \min(V_{30}) > 9345 \wedge \\ \min(V_{113}) \leq 5315 \wedge \\ \langle DR \rangle (\min(V_{97}) \leq 12820 \wedge \langle DR \rangle \min(V_{79}) \\ \geq 6714) \Rightarrow C_{15}, \end{array}$$

and its corresponding path is shown in Figure 8. The two propositional rules for C_{15} are responsible for the class precision and specificity of 50.0% and 14.3%, respectively, whereas the RCC5 rule for C_{15} displays a precision and a specificity of 77.0% (equal to its confidence) and 49.0%, respectively.

Finally, qualitative remarks on the behavior of the trained models can be made. Figure 9 shows a visual depiction of the classification results achieved by the two trees in the full Indian Pines scene. The qualitative comparison shows that some of the areas in the image are classified with less confusion by the spatial RCC5 tree with respect to the propositional, single-pixel one, while the opposite does not hold. From Figure 9(c),(d), it can be observed that many misclassifications occur at the edges of the regions, delimiting pixels of the same class. This could be due to several facts, for example, pixel mixing, which may hinder the predictive power of crisp scalar conditions.

Table 4. Confusion matrices on Indian Pines for pure single-pixel approach (top) and pure RCC5 approach (bottom). Each line shows the confusion of samples belonging to a single class, as well as the number of samples and the recall attained for that class. For the two models, precision and overall accuracy values are also shown. For each class, the highest recall and precision across the two models is highlighted, as well as the highest overall accuracy.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	tot.	rec (%)
(a) single-pixel																		
C1	0	0	0	0	0	0	0	45	0	0	1	0	0	0	0	0	46	0
C2	0	484	41	32	0	15	0	1	0	0	828	27	0	0	0	0	1428	33.9
C3	0	40	74	12	0	0	0	0	0	0	607	19	1	0	0	0	753	9.8
C4	0	34	5	95	0	11	0	1	0	0	72	15	0	0	4	0	237	40.1
C5	0	6	0	3	248	131	0	13	0	0	3	1	0	47	1	0	453	54.8
C6	0	0	0	0	3	718	0	0	0	0	3	0	0	4	2	0	730	98.4
C7	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0	28	0
C8	0	24	0	0	0	2	0	448	0	0	1	0	0	0	3	0	478	93.7
C9	0	0	0	0	0	16	0	0	0	0	0	0	0	0	4	0	20	0
C10	0	1	3	2	0	17	0	2	0	0	935	2	0	0	0	0	962	0
C11	0	71	53	2	0	39	0	3	0	0	2216	6	0	0	2	0	2392	92.6
C12	0	27	40	31	0	9	0	0	0	0	379	98	6	0	0	0	590	16.6
C13	0	0	0	0	0	18	0	0	0	0	0	0	178	0	9	0	205	86.8
C14	0	0	0	0	45	12	0	0	0	0	0	0	1	1187	20	0	1265	93.8
C15	0	0	0	9	8	157	0	6	0	0	1	0	26	62	45	0	314	14.3
C16	0	8	0	1	0	11	0	0	0	0	71	2	0	0	0	0	93	0
prec (%)	-	69.6	34.3	50.8	81.6	62.1	-	81.9	-	-	43.3	57.6	84	91.3	50	-	acc = 58.0	
# rules	0	5	3	4	2	1	0	1	0	0	3	3	1	3	2	0	rt = 28	
(b) RCC5																		
C1	0	0	0	0	0	0	0	39	0	0	7	0	0	0	0	0	46	0
C2	0	620	12	0	0	0	0	0	1	0	781	14	0	0	0	0	1428	43.4
C3	0	14	138	23	0	0	0	0	0	0	557	21	0	0	0	0	753	18.3
C4	0	54	9	97	0	1	0	0	5	0	43	15	13	0	0	0	237	40.9
C5	0	3	0	4	381	14	0	7	11	0	17	0	2	13	1	0	453	84.1
C6	0	0	0	0	5	711	0	9	0	0	1	0	0	4	0	0	730	97.4
C7	0	0	0	0	0	0	0	27	0	0	0	0	0	0	1	0	28	0
C8	0	0	0	0	0	0	0	476	0	0	0	0	0	0	2	0	478	99.6
C9	0	0	0	0	2	8	0	0	7	0	3	0	0	0	0	0	20	35
C10	0	5	4	0	0	0	0	4	0	0	949	0	0	0	0	0	962	0
C11	0	18	25	4	0	0	0	3	9	0	2322	11	0	0	0	0	2392	97.1
C12	0	37	15	7	0	0	0	2	4	0	319	193	8	0	5	0	590	32.7
C13	0	0	0	0	0	0	0	0	6	0	13	0	174	0	12	0	205	84.9
C14	0	0	0	0	17	17	0	0	0	0	0	0	0	1206	25	0	1265	95.3
C15	0	0	0	0	19	68	0	1	0	0	0	0	9	63	154	0	314	49
C16	0	17	1	0	0	0	0	21	3	0	49	0	2	0	0	0	93	0
prec (%)	-	80.7	67.7	71.8	89.9	86.8	-	80.8	15.2	-	45.9	76	83.6	93.8	77	-	acc = 64.8	
# rules	0	4	4	2	1	2	0	1	1	0	4	3	1	1	1	0	rt = 25	

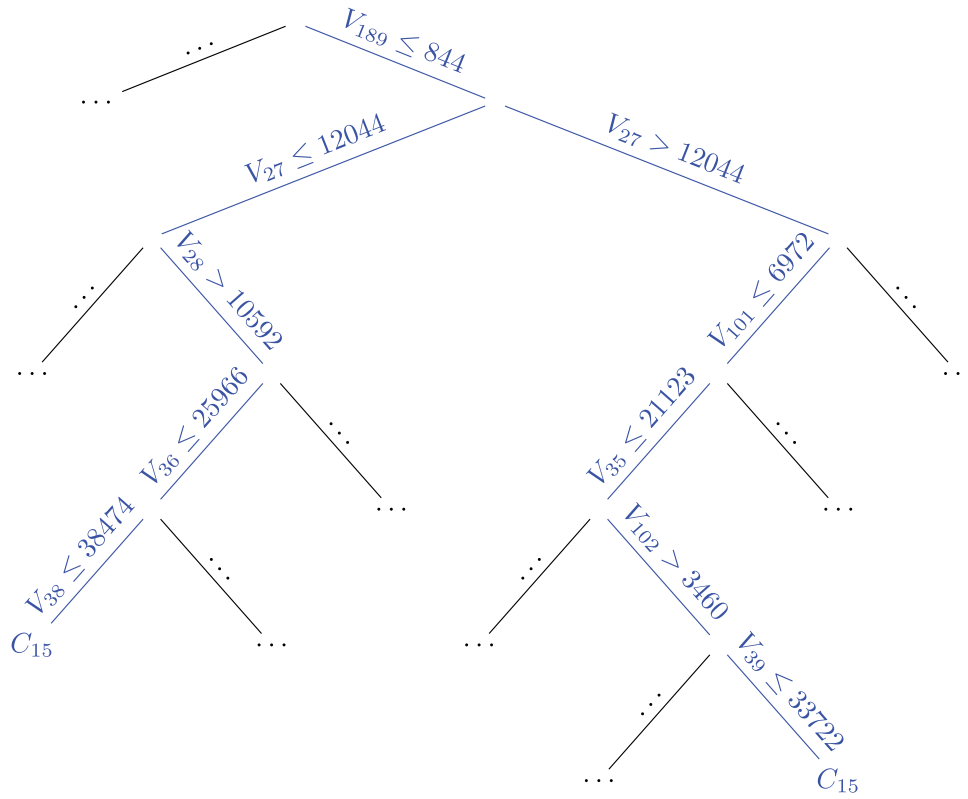


Figure 7. Propositional decision tree trained via the pure single-pixel approach on Indian Pines. While tree comprehends 28 leaves, only the two paths for class C_{15} are shown.

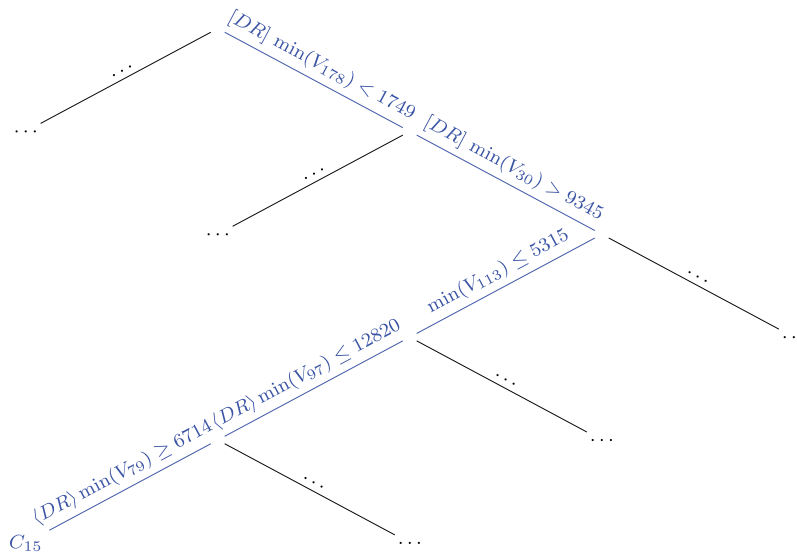


Figure 8. Spatial decision tree trained via the pure RCC5 approach on Indian Pines. While tree comprehends 25 leaves, the single path for class C_{15} is shown.

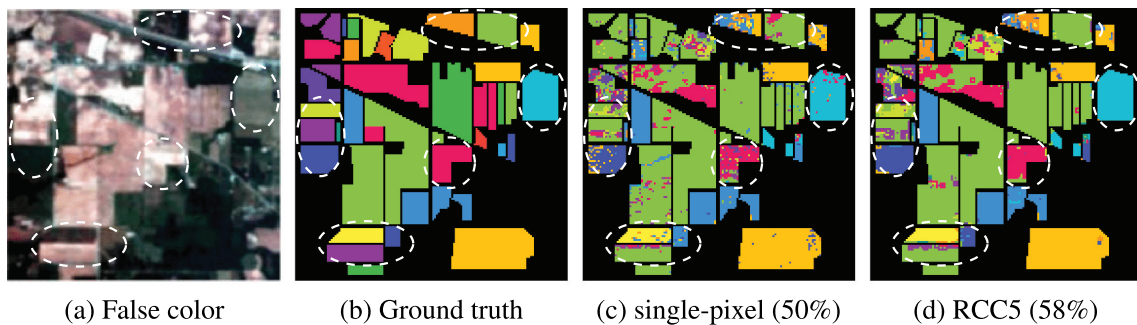


Figure 9. Qualitative result comparison on Indian Pines: false color image (a), ground truth (b), and classification results obtained by the pure single-pixel approach (c) and the pure RCC5 approach (d). The two models scored a κ coefficient equal to 50% and 58%, respectively. A few regions where the RCC5 approach appears to improve over the single-pixel approach are highlighted.

Conclusion

In this article, we proposed a novel technique for knowledge extraction from spatial data. In the context of a general approach that may be called *modal symbolic learning*, we considered a well-known symbolic learning schema, namely decision trees, and enhanced it by substituting propositional logic with a suitable modal spatial logic. This method can capture patterns in the spatial arrangements of objects and, therefore, make more informed decisions toward the classification of images. This work should be looked at from different perspectives. First, from a purely statistical point of view, spatial decision trees proved to have better generalization capability than purely propositional ones and were good in absolute terms for the LCC task and the chosen benchmark datasets. Second, from a foundational standpoint, spatial decision trees are located at the intersection between the symbolic learning theory and the modal logic theory, therefore opening a new field of research in which both well-known and new modal languages can be studied from a new perspective (that is, induction). Third, from a machine learning perspective, spatial symbolic learning can be a serious alternative to functional learning, typically based on neural networks, being able to offer statistically accurate models that can, unlike functional ones, give rise to a logical theory of the phenomena under consideration and therefore be interpreted in a much natural way. In this respect, one should add that it is customary in machine learning to judge a new learning model uniquely from its statistical performances, which is in fact very reductive; being able to extract a theory that can be analyzed, validated, and corrected with background knowledge that can be written in the same language should compensate even the loss of some points in the accuracy of a classification exercise.

One of the downsides of having interpretable decisions as simple as scalar conditions is that their predictive power is rather limited; this reveals when considering that simple conditions are not resilient to simple phenomena such as pixel mixing. A few generalizations of this approach can help mitigate these effects. First, a fuzzy extension of the presented modal decision tree learning framework can be explored; by means of specific, so-called membership functions, this would allow for the truth of a scalar condition (e.g. $\min(V) > \nu$) on a given to be expressed by non-Boolean values (for example, normalized values in $[0, 1]$ expressing a *degree of truth*). Second, spatial decision trees with more complex feature functions can be explored; as shown in (Pagliarini et al., 2022), feature functions such as min and max can be replaced by neural networks specifically trained for feature extraction (e.g. by means of autoencoding), and, while the resulting model loses interpretability,

it potentially becomes more accurate. This approach leads to an instance of *neural-symbolic* computation, which is a trending trade-off with respect to the functional-symbolic dichotomy (Besold et al., 2021). Finally, there is a significant computational limitation to the spatial decision tree learning algorithm presented, which currently requires intensive computation of the properties of all rectangles in an image. Several methods can help reduce the computational load, and arguably the most promising one uses reduced spatial models, in which not all rectangles are considered but only the most informative one, according to a given information measure.

In conclusion, in the recent literature, symbolic learning has been eclipsed by the explosion of functional learning methods and, in particular, neural network-based methods for pattern recognition and data-driven extraction of knowledge; the reasons behind this phenomenon include the better statistical performances and the apparent universality of the latter approaches over the former ones. But one key observation is that symbolic learning has been confined to the language of propositional logic, which has very limited expressive power; the introduction of modal, and in particular spatial, symbolic learning could be a step toward balancing out this duality.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

We acknowledge the support of the INDAM-GNCS project Symbolic and Numerical Analysis of Cyberphysical Systems (code CUP-E53C22001930001), funded by INDAM, and of the FIRD project Symbolic Geometric Learning, funded by the University of Ferrara.

ORCID

G. Pagliarini  <http://orcid.org/0000-0002-8403-3250>
G. Sciacicco  <http://orcid.org/0000-0002-9221-879X>

Data availability statement

The data that support the findings of this study are available from the corresponding author, G. Pagliarini, upon reasonable request.

References

- Aceto, L., Della Monica, D., Goranko, V., Ingólfssdóttir, A., Montanari, A., & Sciacicco, G. (2016). A complete classification of the expressiveness of interval logics of Allen's relations: The general and the dense cases. *Acta*

- Informatica*, 53(3), 207–246. <https://doi.org/10.1007/s00236-015-0231-4>
- Ahmad, M., Khan, A. M., & Hussain, R. (2017). Graph-based spatial-spectral feature learning for hyperspectral image classification. *IET Image Processing*, 11(12), 1310–1316. <https://doi.org/10.1049/iet-ipr.2017.0168>
- Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11), 832–843. <https://doi.org/10.1145/182.358434>
- Audebert, N., Le Saux, B., & Lefèvre, S. (2019). Deep learning for classification of hyperspectral data: A comparative review. *IEEE Geoscience and Remote Sensing Magazine*, 7(2), 159–173. <https://doi.org/10.1109/MGRS.2019.2912563>
- Balbani, P., Condotta, J., & Fariñas Del Cerro, L. (1998). A model for reasoning about bidimensional temporal relations. In *Proceedings of the 6th international conference on principles of knowledge representation and reasoning (KR'98)*, June 2–5, 1998, Trento, Italy (pp. 124–130). Morgan Kaufmann.
- Belgiu, M., Draǵu, L., & Strobl, J. (2014). Quantitative evaluation of variations in rule-based classifications of land cover in urban neighbourhoods using worldview-2 imagery. *The ISPRS Journal of Photogrammetry and Remote Sensing*, 87, 205–215. <https://doi.org/10.1016/j.isprsjprs.2013.11.007>
- Benavoli, A., Corani, G., & Mangili, F. (2016). Should we really use post-hoc tests based on mean-ranks? *The Journal of Machine Learning Research*, 17(1), 152–161. <https://doi.org/10.5555/2946645.2946650>
- Berhane, T., Lane, C., Wu, Q., Autrey, B., Anenkhonov, O., Chepinoga, V., & Liu, H. (2018). Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory. *Remote Sensing*, 10(580), 1–26. <https://doi.org/10.3390/rs10040580>
- Besold, T. R., d'Avila Garcez, A. S., Bader, S., Bowman, H., Domingos, P. M., Hitzler, P., ... Zaverucha, G. (2021). Neural-symbolic learning and reasoning: A survey and interpretation. In P. Hitzler & M.K. Sarker (Eds.), *Neuro-symbolic artificial intelligence: The state of the art* (Vol. 342, pp. 1–51). IOS Press. <https://doi.org/10.3233/FAIA210348>
- Blackburn, P., de Rijke, M., & Venema, Y. (2001). *Modal logic*. Cambridge University Press.
- Blockeel, H., & De Raedt, L. (1998). Top-down induction of first-order logical decision trees. *Artificial Intelligence*, 101(1), 285–297. [https://doi.org/10.1016/S0004-3702\(98\)00034-4](https://doi.org/10.1016/S0004-3702(98)00034-4)
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bresolin, D., Della Monica, D., Montanari, A., Sala, P., & Sciavicco, G. (2014). Interval temporal logics over strongly discrete linear orders: Expressiveness and complexity. *Theoretical Computer Science*, 560, 269–291. <https://doi.org/10.1016/j.tcs.2014.03.033>
- Bresolin, D., Della Monica, D., Montanari, A., & Sciavicco, G. (2014). The light side of interval temporal logic: The Bernays-Schönfinkel fragment of CDT. *Annals of Mathematics and Artificial Intelligence*, 71(1–3), 11–39. <https://doi.org/10.1007/s10472-013-9337-y>
- Bresolin, D., Kurucz, A., Muñoz-Velasco, E., Ryzhikov, V., Sciavicco, G., & Zakharyashev, M. (2017). Horn fragments of the Halpern-Shoham interval temporal logic. *ACM Transactions on Computational Logic*, 18(3), 22:1–22:39. <https://doi.org/10.1145/3105909>
- Bresolin, D., Sala, P., Della Monica, D., Montanari, A., & Sciavicco, G. (2010). A decidable spatial generalization of metric interval temporal logic. In *Proceedings of the 17th international symposium on temporal representation and reasoning (TIME 2010)*, 6–8 September 2010, Paris, France (pp. 95–102).
- Cao, X., Zhou, F., Xu, L., Meng, D., Xu, Z., & Paisley, J. (2018). Hyperspectral image classification with Markov random fields and a convolutional neural network. *IEEE Transactions on Image Processing*, 27(5), 2354–2367. <https://doi.org/10.1109/TIP.2018.2799324>
- Coccagna, M., Manzella, F., Mazzacane, S., Pagliarini, G., & Sciavicco, G. (2022). Statistical and symbolic neuroaesthetics rules extraction from EEG signals. In *Proceedings of the 9th international work-conference on the interplay between natural and artificial computation (IWINAC 2022)*, May 31 - June 3, 2022, Puerto de la Cruz, Tenerife, Spain (Vol. 13258, pp. 536–546). Springer.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Cohn, A. (1995). A hierarchical representation of qualitative shape based on connection and convexity. In *Proceedings of the international conference on spatial information theory: A theoretical basis for GIS (COSIT)*, Semmering, Austria, 988, pp. 311–326).
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7(1), 1–30. <https://doi.org/10.5555/1248547.1248548>
- Deng, H. (2019). Interpreting tree ensembles with inTrees. *International Journal of Data Science and Analytics*, 7(4), 277–289. <https://doi.org/10.1007/s41060-018-0144-8>
- Dubba, K., Cohn, A., Hogg, D., Bhatt, M., & Dylla, F. (2015). Learning relational event models from video. *Journal of Artificial Intelligence Research*, 53, 41–90. <https://doi.org/10.1613/jair.4395>
- Egenhofer, M., Sharma, J., & Mark, D. (1993). A critical comparison of the 4-intersection and 9-intersection models for spatial relations: Formal analysis. In *Proceedings of the 11th international symposium on computer-assisted cartography (Auto-Carto)-11*, October 30–November 1, 1993, Minneapolis, Minnesota (pp. 1–13).
- Frank, A. (1996). Qualitative spatial reasoning: Cardinal directions as an example. *International Journal of Geographical Information Science*, 10(3), 269–290. <https://doi.org/10.1080/02693799608902079>
- Freksa, C. (1992). Using orientation information for qualitative spatial reasoning. In *Theories and methods of spatio-temporal reasoning in geographic space* (Vol. 639, pp. 162–178). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-55966-3_10
- Friedl, M., & Brodley, C. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, 61(3), 399–409. [https://doi.org/10.1016/S0034-4257\(97\)00049-7](https://doi.org/10.1016/S0034-4257(97)00049-7)
- Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3), 2 (3). <https://doi.org/10.1214/07-AOAS148>
- Goel, P., Prasher, S., Patel, R., Landry, J., Bonnell, R., & Viau, A. (2003). Classification of hyperspectral data by decision trees and artificial neural networks to identify weed stress and nitrogen status of corn. *Computers and Electronics in Agriculture*, 39(2), 67–93. [https://doi.org/10.1016/S0168-1699\(03\)00020-6](https://doi.org/10.1016/S0168-1699(03)00020-6)
- Halpern, J., & Shoham, Y. (1991). A propositional modal logic of time intervals. *Journal of the ACM*, 38(4), 935–962. <https://doi.org/10.1145/115234.115351>
- Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd international conference on document analysis and*

- recognition (icdar), August 14-16, 1995, Montreal, Quebec, Canada (pp. 278–282).
- Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., & Chanussot, J. (2022). Spectralformer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–15. <https://doi.org/10.1109/TGRS.2022.3172371>
- Hu, W., Huang, Y., Li, W., Zhang, F., & Li, H. (2015). Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors*, 2015, 258619:1–258619:12. <https://doi.org/10.1155/2015/258619>
- Hyafil, L., & Rivest, R. L. (1976). Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1), 15–17. [https://doi.org/10.1016/0020-0190\(76\)90095-8](https://doi.org/10.1016/0020-0190(76)90095-8)
- Jiang, J., Ma, J., Wang, Z., Chen, C., & Liu, X. (2019). Hyperspectral image classification in the presence of noisy labels. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2), 851–865. <https://doi.org/10.1109/TGRS.2018.2861992>
- Jiang, Z., Shekhar, S., Mohan, P., Knight, J., & Corcoran, J. (2012). Learning spatial decision tree for geographical classification: A summary of results. In *Proceedings of the 12th international conference on advances in geographic information systems*, November 6-9, 2012, Redondo Beach, California, USA (p. 390–393). ACM.
- Kartikeyan, B., Majumder, K., & Dasgupta, A. (1995). An expert system for land cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, 33(1), 58–66. <https://doi.org/10.1109/36.368222>
- Kucharczyk, M., Hay, G. J., Ghaffarian, S., & Hugenholtz, C. H. (2020). Geographic object-based image analysis: A primer and future directions. *Remote Sensing*, 12(12), 2012. <https://doi.org/10.3390/rs12122012>
- Kulkarni, A., & Shrestha, A. (2017). Multispectral image analysis using decision trees. *International Journal of Advanced Computer Science & Applications*, 8(6), 11–18. <https://doi.org/10.14569/IJACSA.2017.080602>
- Lee, H., & Kwon, H. (2017). Going deeper with contextual CNN for hyperspectral image classification. *IEEE Transactions on Image Processing*, 26(10), 4843–4855. <https://doi.org/10.1109/TIP.2017.2725580>
- Liaw, A., & Wiener, M. (2002). Classification and regression by RandomForest. *R News*, 2(3), 18–22.
- Li, Y., Zhang, H., & Shen, Q. (2017). Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sensing*, 9(1), 67. <https://doi.org/10.3390/rs9010067>
- Lutz, C., & Wolter, F. (2006). Modal logics of topological relations. *Logical Methods in Computer Science*, 2(2), 1–41. [https://doi.org/10.2168/LMCS-2\(2:5\)2006](https://doi.org/10.2168/LMCS-2(2:5)2006)
- Makantasis, K., Karantzas, K., Doulamis, A. D., & Doulamis, N. D. (2015). Deep supervised learning for hyperspectral data classification through convolutional neural networks. In *Proceedings of the international geoscience and remote sensing symposium (IGARSS)*, July 26-31, 2015, Milan, Italy (pp. 4959–4962). IEEE.
- Malerba, D., Ceci, M., & Appice, A. (2005). Mining model trees from spatial data. In *Proceedings of the 9th european conference on principles and practice of knowledge discovery in databases*, October 3-7, 2005, Porto, Portugal (pp. 169–180). Springer.
- Manzella, F., Pagliarini, G., Sciacicco, G., & Stan, I. (2023). The voice of COVID-19: Breath and cough recording classification with temporal decision trees and random forests. *Artificial Intelligence in Medicine*, 137, 102486. <https://doi.org/10.1016/j.artmed.2022.102486>
- Marcinkowski, J., & Michaliszyn, J. (2014). The undecidability of the logic of subintervals. *Fundamenta Informaticae*, 131(2), 217–240. <https://doi.org/10.3233/FI-2014-1011>
- Marx, M., & Reynolds, M. (1999). Undecidability of compass logic. *Journal of Logic and Computation*, 9(6), 897–914. <https://doi.org/10.1093/logcom/9.6.897>
- Meinshausen, N. (2010). Node harvest. *The Annals of Applied Statistics*, 4(4). <https://doi.org/10.1214/10-AOAS367>
- Montanari, A., Pratt-Hartmann, I., & Sala, P. (2010). Decidability of the logics of the reflexive sub-interval and super-interval relations over finite linear orders. In *Proceedings of the 17th international symposium on temporal representation and reasoning (TIME 2016)*, 6-8 September 2010, Paris, France (pp. 27–34).
- Montanari, A., Puppis, G., & Sala, P. (2009). A decidable spatial logic with cone-shaped cardinal directions. In *Proceedings of the 18th annual conference of the european association for computer science logic (CSL 2009)*, September 7-11, 2009, Coimbra, Portugal (Vol. 5771, pp. 394–408). Springer.
- Montanari, A., Sciacicco, G., & Vitacolonna, N. (2002). Decidability of interval temporal logics over split-frames via granularity. In *Proceedings of the 8th european conference on logics in artificial intelligence (JELIA 2002)*, September, 23-26, Cosenza, Italy (Vol. 2424, pp. 259–270). Springer.
- Monteiro, S. T., & Murphy, R. J. (2011). Embedded feature selection of hyperspectral bands with boosted decision trees. In *Proceedings of the international geoscience and remote sensing symposium (IGARSS)*, July 24-29, 2011, Vancouver, BC, Canada (pp. 2361–2364). IEEE.
- Morales Nicolás, A., Navarrete, I., & Sciacicco, G. (2007). A new modal logic for reasoning about space: Spatial propositional neighborhood logic. *Annals of Mathematics and Artificial Intelligence*, 51(1), 1–25. <https://doi.org/10.1007/s10472-007-9083-0>
- Mou, L., Ghamisi, P., & Zhu, X. (2017). Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7), 3639–3655. <https://doi.org/10.1109/TGRS.2016.2636241>
- Muñoz-Velasco, E., Pelegrín-García, M., Sala, P., Sciacicco, G., & Stan, I. (2019). On coarser interval temporal logics. *Artificial Intelligence*, 266, 1–26. <https://doi.org/10.1016/j.artint.2018.09.001>
- Navarrete, I., Morales Nicolás, A., Sciacicco, G., & Cárdenas-Viedma, M. A. (2013). Spatial reasoning with rectangular cardinal relations: The convex tractable sub-algebra. *Annals of Mathematics and Artificial Intelligence*, 67(1), 31–70. <https://doi.org/10.1007/s10472-012-9327-5>
- Pagliarini, G., Manzella, F., Sciacicco, G., & Stan, I. E. (2023). *ModalDecisionTrees.Jl: Interpretable Models for Native Time-Series & Image Classification*. <https://github.com/aclai-lab/ModalDecisionTrees.jl>
- Pagliarini, G., Scaboro, S., Serra, G., Sciacicco, G., & Stan, I. E. (2022). Neural-symbolic temporal decision trees for multivariate time series classification. In *29th international symposium on temporal representation and reasoning (TIME 2022)*, 7-9 November, 2022, Virtual (in press)
- Pal, M., & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment*, 86(4), 554–565. [https://doi.org/10.1016/S0034-4257\(03\)00132-9](https://doi.org/10.1016/S0034-4257(03)00132-9)
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A.,

- Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://doi.org/10.5555/1953048.2078195>
- Phiri, D., Simwanda, M., Nyirenda, V., Murayama, Y., & Ranagalage, M. (2020). Decision tree algorithms for developing rulesets for object-based land cover classification. *ISPRS International Journal of Geo-Information*, 9 (5), 329. <https://doi.org/10.3390/ijgi9050329>
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106. <https://doi.org/10.1007/BF00116251>
- Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.
- Randell, D., Cui, Z., & Cohn, A. (1992). A spatial logic based on regions and connection. In *Proceedings of the 3rd international conference on principles of knowledge representation and reasoning (KR'92)*, October 25–29, 1992, Cambridge, Massachussets (pp. 165–176).
- Renz, J., & Nebel, B. (2007). Qualitative spatial reasoning using constraint calculi. In *Handbook of spatial logics* (pp. 161–215). Springer Netherlands. https://doi.org/10.1007/978-1-4020-5587-4_4
- Roy, S., Krishna, G., Dubey, S., & Chaudhuri, B. (2020). Hybridsn: Exploring 3-d-2-d CNN feature hierarchy for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 17(2), 277–281. <https://doi.org/10.1109/LGRS.2019.2918719>
- Sadeghi, B. (2013). DecisionTree.jl. <https://github.com/JuliaAI/DecisionTree.jl>
- Santara, A., Mani, K., Hatwar, P., Singh, A., Garg, A., Padia, K., & Mitra, P. (2017). BASS net: Bandadaptive spectral-spatial feature learning neural network for hyperspectral image classification. *IEEE Transactions on Geoscience & Remote Sensing*, 55(9), 5293–5301.
- Sciavicco, G., & Stan, I. (2020). Knowledge extraction with interval temporal logic decision trees. In E. Muñoz-Velasco, A. Ozaki & M. Theobald (Eds.), *Proceedings Of the 27th international symposium on temporal representation and reasoning (TIME 2020)* (Vol. 178, pp. 9:1–9:16). Bolzano, Italy: Schloss Dagstuhl.
- Skiadopoulos, S., & Koubarakis, M. (2004). Composing cardinal direction relations. *Artificial Intelligence*, 152(2), 143–171. [https://doi.org/10.1016/S0004-3702\(03\)00137-1](https://doi.org/10.1016/S0004-3702(03)00137-1)
- Skiadopoulos, S., Sarkas, N., Sellis, T., & Koubarakis, M. (2007). A family of directional relation models for extended objects. *IEEE Transactions on Knowledge and Data Engineering*, 19(8), 1116–1130. <https://doi.org/10.1109/TKDE.2007.1046>
- Tüysüzoğlu, G., Birant, D., & Kiranoğlu, V. (2022). Temporal bagging: A new method for time-based ensemble learning. *Turkish Journal of Electrical Engineering and Computer Sciences*, 30(1), 279–294. <https://doi.org/10.3906/elk-2011-41>
- Walega, P. A., & Zawidzki, M. (2019). A modal logic for subject-oriented spatial reasoning. In J. Gamper, S. Pinchinat & G. Sciavicco (Eds.), *Proceedings Of the 26th international symposium on temporal representation and reasoning (TIME'19)* (Vol. 147, pp. 4:1–4:22). Málaga, Spain: Schloss Dagstuhl.
- Witten, I. H., Frank, E., & Hall, M. A. (2017). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.
- Xu, M., Watanachaturaporn, P., Varshney, P., & Arora, M. (2005). Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, 97(3), 322–336. <https://doi.org/10.1016/j.rse.2005.05.008>
- Zhang, Q., & Wang, J. (2003). A rule-based urban land use inferring method for fine-resolution multispectral imagery. *Canadian Journal of Remote Sensing*, 29(1), 1–13. <https://doi.org/10.5589/m02-075>