Array 30 (2026) 100893

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array

Beyond tabular data: Feature extraction and selection for high-dimensional data

P. Cavina [a], F. Manzella [b], G. Pagliarini [b], G. Sciavicco [b,*], I.E. Stan [c]

[a] Department of Mathematics, Computer Science and Physics, University of Udine, Udine, 33100, Italy
[b] Department of Mathematics and Computer Science, University of Ferrara, Ferrara, 44121, Italy
[c] Department of Informatics, Systems, and Communications, University of Milano-Bicocca, Milano, 20126, Italy

ARTICLE INFO

Keywords:
Feature selection
Feature extraction
Dimensional data

ABSTRACT

The advent of big data has catalyzed the development of numerous data-driven applications across various fields. Essential to the efficacy of these applications is the ability to process and analyze high-dimensional datasets, such as time series, images, and beyond. This research focuses on advancing the methodologies for feature extraction and selection tailored to the unique challenges presented by such data. While feature extraction and selection have been considerably investigated for tabular data, their application to high-dimensional data still needs to be examined. In this paper, we provide a protocol to both reduce dimensionality and retain critical information from high-dimensional datasets necessary for modeling and prediction tasks—specifically designed for, although not limited to, symbolic approaches—using, at a lower level, state-of-the-art feature selection algorithms for tabular data in a systematic way. We demonstrate the effectiveness of our approaches using real-world examples and evaluate their performance using appropriate, objective metrics, and we provide a complete open-source package as part of a long-term project for symbolic non-tabular data representation and learning. Our results contribute to the ongoing discourse on data science and machine learning practices, with particular attention to symbolic approaches.

1. Introduction

Structured and non-structured data. In the current era of data proliferation, the vast majority of data is unstructured, meaning it does not conform to traditional tabular formats and presents complexities in analysis using conventional data analytics methods. This category encompasses time series, images, and various other forms of qualitative data characterized by their need for subjective interpretation. According to [1], a staggering 95% of all existing data is non-tabular, emphasizing the pressing demand to adeptly handle and analyze such data in today's big data landscape. Conversely, structured data's organization within preconceived schema renders it straightforward to process and plays a pivotal role across numerous fields, enhancing the effectiveness of tasks in financial analysis and customer relationship management, to name but two.

Selection of features. At the core of data analysis are processes such as inspection, cleansing, transformation, and abstraction, all converging towards the revelation of valuable insights that support informed decision-making. At this intersection lie *feature extraction* and *feature selection*, techniques that are fundamental to preparing a dataset for subsequent analysis. Feature extraction seeks to distill raw data into a more manageable and insightful lower-dimensional space without sacrificing its inherent complexities. Complementarily, feature selection aims to pinpoint the most informative attributes by pruning away redundancy and noise. Feature extraction and selection, however, should not be solely in the light of a subsequent automatic knowledge extraction phase; instead, these processes should be evaluated also by their ability of highlighting *per se* the relevant information hidden in data.

Feature extraction employs an array of techniques, ranging from the simple, such as scaling and normalization, to the complex, like principal component analysis (PCA) or singular value decomposition (SVD) [2], and it can be *univariate*, when variables are modified independently or *multivariate*, when several variables are combined into other variables.

Feature selection methodologies [3–14], on the other hand, range from *filter-based*, to *wrapper-based*, *embedded*, and *hybrid* approaches. Filter-based methods are independent of the learning models and base their estimate of feature importance on heuristic or statistical ranking criteria. Two orthogonal dimensions characterize filters: *univariate/multivariate* approaches and *supervised/unsupervised* settings. Univariate filters rank each feature independently and thus disregard any form

* Corresponding author.
E-mail address: scvgdu@unife.it (G. Sciavicco).

https://doi.org/10.1016/j.array.2026.100893
Received 29 June 2025; Received in revised form 12 May 2026; Accepted 12 May 2026
Available online 16 May 2026

of inter-feature correlation. Multivariate filters, on the contrary, rank multiple features in batches and can account for interdependencies between features [15]. Supervised filters estimate the relevance of a feature (group) concerning a target variable, while unsupervised ones select features only according to their nature and characteristics [16]. Unsupervised methods include filters based on the variance, entropy, and Laplacian score [17], while supervised ones enclose correlation/covariance, entropy gain, supervised Laplacian score, hypothesis test(s), and discriminating power [18], depending on their statistical principle. Wrapper methods, unlike filters, use a learning model to evaluate the performance of different feature subsets: iterate a search process until an optimal result, or some stopping condition, is reached, where the search strategies can be random, sequential, or heuristic [19]. Embedded methods integrate the feature selection process into the learning model. Finally, hybrid methods are combinations of filters and wrappers that take advantage of both techniques: the filter reduces the feature set to a good enough subset, and the wrapper maximizes performance [15].

Dimensional data. A prevalent form of non-tabular data is *dimensional data*, where each instance comprises an n-uple of real functions set against a discrete d-dimensional grid, like $\mathbb{N}^d$. This type encompasses a broad spectrum of real-world data, including temporal (when $d = 1$), spatial (for $d = 2$ or $d = 3$), and video (with $d = 3$) datasets, not excluding scalar data (where $d = 0$). Traditionally, machine learning endeavors in these realms have been dominated by sub-symbolic approaches, particularly neural networks, which present an abundant and sprawling literature that is too vast for thorough review within the scope of this paper. Of late, there has been a burgeoning interest in symbolic learning, especially applied to temporal and spatial datasets. Symbolic techniques have made strides for tasks such as temporal rule extraction [20] and the classification of time series—employing methods that range from decision trees [21] to shapelet-based algorithms [22] and logic-based classifiers designed to distill point-based temporal logic formulas [23], while at the same time applying symbolic methods in the spatial domain, especially propositional decision trees, has also yielded promising results [24].

A novel symbolic methodology, applicable across all types of dimensional data (on top of other unstructured data), that has recently gathered momentum is *modal symbolic learning* [25]. In a nutshell, this approach supplants traditional propositional logic with modal logic to represent and reason with patterns within d-dimensional data, by focusing on the analysis of d-dimensional hyperintervals and their qualitative relationships, and capturing their relationships with suitably expressive modal logics, effectively offering a fresh perspective on modern machine learning. Despite its recent inception, the theoretical underpinnings of modal symbolic learning are already a subject of active research [26–28], and implementations such as modal decision trees forests have validated their utility in various cases [29–31].

Our contribution. In their d-dimensional version, modal symbolic techniques naturally inspire a protocol for feature extraction and selection based on the idea of d-dimensional *hyperintervals*, which can be seen as the obvious generalization of 1-dimensional intervals to the d dimensions case. On the one side, the current approaches to learning from dimensional data produced a number proposals for feature extraction functions, ranging from obvious basic statistical measures (e.g., *minimum*, *maximum*, and *mean*) to more complex functions, both general-purpose, like *Catch22* [32], and specific, like *Mel Frequency Cepstral Coefficients (MFCC)* [33], for audio and EEG data. On the other hand, hyperintervals allow one to apply any such function to a portion of each instance, preserving their relative dimensional relationships. Thus, a natural approach to extraction and selection of features from dimensional data starts from applying a range of such functions to range of hyperintervals. As it appears immediately clear, however, this causes an information explosions problem, which can be summarized in the following question: *which feature functions applied to which variables in which hyperintervals contain the most information?* While it could be argued that the answer to this question lies in standard selection techniques, the fact is that even with relatively small d-dimensional datasets the problem may become quickly unaffordable from the computational point of view. As a consequence, a first exploratory selection step must be limited to *univariate extraction, selection, filter-based* tools. Although this model for data representation is not uncommon, it has never been used to encompass data of different nature under the same feature selection methodology. The literature shows that attempts to taxonomize feature selection methods [34] often create different classes of methodologies for different data types such as time series [35–38] or images [39,40]. This is particularly clear when analyzing the documentation of the most popular software tools for feature selection such as *scikit-learn* [41] (written in the language Python) and *caret* [42] (written in the language R).

Our contribution aims at filling in these gaps. We design, implement, and test a clearly designed protocol that includes fully parameterizable methods applicable to any set of feature extraction functions (including sub-symbolic approaches, to some extent) as well as a reliable statistical evaluation of the results; our code is included in *Sole.jl* [43], an open-source framework dedicated to symbolic methods for data analysis, learning, and model analysis. To demonstrate the effectiveness of our approach, we consider six very complex datasets, three in the temporal case and three in the spatial case. In the temporal case we tackle the problem of extracting and selecting features from *human electroencephalogram signals*; our technique will allow us to identify several useful elements, including the most informative electrodes, the most informative feature functions, the most informative brain wave frequencies, and the most informative intervals of time (across the trials) for the considered problems. Similarly, in the spatial case, we shall extract the most informative (rectangular) areas, the most informative light frequencies, and the most informative feature extraction functions in three supervised land cover classification problems from *satellite images*. In both cases, the number of combinations among variables, feature extraction functions, and hyperintervals exceed the several thousands, and an informed choice for their selection is of uttermost importance for any successive task. All data are publicly and freely available.

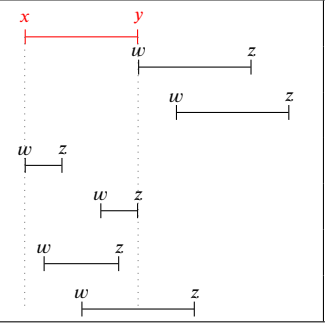
Limitations. Our approach is not currently designed to cover the case of wrapper-based and embedded selection methods. The main cause is intrinsic to the nature of filters being separable from the learning phase, so that data can be freely manipulated to extract the necessary information for the selection. Symbolic learning from dimensional data is still being explored, and the current technological level seems to suggest that wrapper-based or embedded selection is still not possible and/or computationally affordable for dimensional data just yet; however, our framework can be considered as a first step in this direction where such possibilities could, at least in principle, be studied.

Organization. This paper is organized as follows. Section 2 delves into the theoretical underpinnings of our methodology, emphasizing the symbolic perspective that distinguishes our work from conventional data analysis paradigms. Following this, Section 3 presents the development and application of our open-source software package to facilitate the practical implementation of our symbolic analysis techniques. Then, Section 4 comprehensively examines the empirical findings derived from applying our methodologies, highlighting the challenges encountered and the insights gained, before concluding.

2. Dimensional data

Preliminaries. Given a predefined set of *variables* (or *variable names*) $\mathcal{V} = \{V_1, \ldots, V_n\}$, a (0-*dimensional*) *tabular instance* is an object of the form $I = (v_1, \ldots, v_n)$, where each *value* $v_i \in \mathbb{R}$ refers to the variable V_i. A (0-*dimensional*) *tabular dataset* is thus a set $\mathcal{I} = \{I_1, \ldots, I_m\}$ of tabular instances. Additionally, tabular datasets may be *labeled*, wherein each

Table 1
Allen's interval relations and $\mathcal{HS}^1$ modalities. Equality (=) is not displayed.

$\mathcal{HS}$ modality	1-dimensional model definition			Example
$\langle A \rangle$ (after)	$[x, y]R_A[w, z]$	$\Leftrightarrow$	$y = w$	
$\langle L \rangle$ (later)	$[x, y]R_L[w, z]$	$\Leftrightarrow$	$y < w$	
$\langle B \rangle$ (begins)	$[x, y]R_B[w, z]$	$\Leftrightarrow$	$x = w \wedge z < y$	
$\langle E \rangle$ (ends)	$[x, y]R_E[w, z]$	$\Leftrightarrow$	$y = z \wedge x < w$	
$\langle D \rangle$ (during)	$[x, y]R_D[w, z]$	$\Leftrightarrow$	$x < w \wedge z < y$	
$\langle O \rangle$ (overlaps)	$[x, y]R_O[w, z]$	$\Leftrightarrow$	$x < w < y < z$	

instance I corresponds to a unique *label* L from a predefined set $\mathcal{L} = \{L_1, \ldots, L_k\}$.

Commonly, tabular datasets are characterized by scalar values and are frequently utilized within the realms of data science and machine learning. It is important to note that categorical variables can be encoded as scalar values using techniques such as one-hot encoding. However, tabular datasets represent only a specific subtype of the broader category known as dimensional datasets. A *d-dimensional instance* can be described as an n-tuple of functions $I = (v_1, \ldots, v_n)$, where each v_i is a *function*

$$v_i : \mathbb{N}^d \to \mathbb{R}.$$

Correspondingly, a *d-dimensional dataset* is a collection $\mathcal{I} = \{I_1, \ldots, I_m\}$ of d-dimensional instances; similarly to tabular datasets, d-dimensional datasets can be labeled. This conceptual framework extends the application scope of tabular datasets, which can be now seen as a special case of d-dimensional ones (with $d = 0$), enabling the encapsulation of time series data ($d = 1$), images ($d = 2$), and video data ($d = 3$), among others. Our discussion hereafter will focus specifically on *finite* datasets such that each dimension of the dimensional space $\mathbb{N}^d$ contains precisely N distinct points.

Extracting information from d-dimensional data is achieved through the employment of a pre-established suite of *feature extraction functions* $\mathcal{F} = \{f_1, \ldots, f_l\}$. Each function f_i within this set is abstractly defined as

$$f_i : \bigcup_{1 \le j_1, \ldots, j_d \le N} \mathbb{R}^{j_1} \times \cdots \times \mathbb{R}^{j_d} \to \mathbb{R}.$$

Each abstractly defined function represents, in fact, a set of functions; for example, if f_i is the *generalized mean* of a set of values, then one element of f_i is the *mean of the values in a 3 × 2 rectangle*. Thus, each abstract function f_i is concretized into a several ones, across all dimensions and possible domains. In general, the set of abstract functions $\mathcal{F}$ may encompass anything from simple and intuitive functions (such as the mean, the minimum, the maximum, and so on) to very complex ones, and even implicitly defined ones (such as a neural network). Applying an abstract function $f \in \mathcal{F}$ to a variable $V \in \mathcal{V}$ results into a *feature* $f(V)$. This approach facilitates the transformation of high-dimensional data into a form amenable to analysis by encapsulating salient information within the extracted features.

It is of notice that extracting functions as features is a common practice in dimensional data analysis; our contribution is the systematization of such a practice in the form of a general framework. This allows us, *inter alia*, to include the possibility of using well-known and accepted feature extraction techniques that are specific for each case as particular cases of a general approach.

Symbolic learning from dimensional data. *Modal symbolic learning* [25] is based on the idea that non-tabular data can be represented with (a suitable) modal logic. There are several possible modal languages that may represent dimensional data; each of them may, in principle, drive a specific feature extraction and selection protocol. Towards a uniform treatment, however, we introduce the logic $\mathcal{HS}^d$, which is the d-dimensional generalization of Halpern and Shoham's modal logic of time intervals $\mathcal{HS}$ [44], and follows the ideas introduced by Balbiani with his Rectangle Algebra [45].

Consider a d-dimensional grid, $\mathbb{D}^d$, where $\mathbb{D} = \langle \{1, \ldots, N\}, \le \rangle$ is a segment of the natural numbers up to some cardinality N. Within this space, we define a *hyperinterval* as a set of d pairs of natural numbers that outline a subregion or 'block' in our d-dimensional space, that is

$$H = ([x_1, y_1], \ldots, [x_d, y_d]) \subseteq \mathbb{D}^d,$$

where each pair $[x_i, y_i]$ defines the bounds of the interval along the ith dimension, with $1 \le x_i \le y_i \le N$ for each dimension i such that $1 \le i \le d$. When $x_i = 1$ and $y_i = N$ across all dimensions, the hyperinterval H encompasses the full extent of $\mathbb{D}^d$. To say that a *point belongs to a hyperinterval* means that the coordinates of the point fall within the respective dimensional intervals of the hyperinterval. Formally, a point $(\pi_1, \ldots, \pi_d)$ is an element of the hyperinterval H, denoted

$$(\pi_1, \ldots, \pi_d) \in H,$$

if and only if $x_i \le \pi_i \le y_i$ holds true across all dimensions i. This representation of d-dimensional datasets allows us to navigate and analyze the entities and their relationships within the data, and provides a structured approach to extracting meaningful insights that might be encoded in the geometry and relative positioning of data points within the multi-dimensional space.

Building on the seminal work of Allen [46], who defined thirteen temporal *interval relations* that succinctly describe the possible relative positions of pairs of intervals on a linear timeline, we extend these concepts to d-dimensional spaces. Allen's interval relations provide a qualitative vocabulary for how intervals can be juxtaposed, such as meeting, overlapping, or preceding one another. As shown in Table 1, each relation R_X, with $X \in \mathcal{X} = \{A, L, B, E, D, O\}$ is associated with a modal operator $\langle X \rangle$; moreover, each relation R_X has a corresponding inverse one $R_{\overline{X}}$. In order to transpose Allen's interval relations into the realm of d-dimensional data, we have to consider hyperintervals. The relationship between two hyperintervals H and H' is denoted as $R_{X_1 \ldots X_d}$, where each $X_i \in \mathcal{X} \cup \{=\}$ represents one of Allen's original interval relations or it is the equality. Just as in the 1-dimensional case the relation $R_=$ is discarded as it does not give rise to a modal operator, in the d-dimensional one the relation $R_{=,=,\ldots,=}$ is excluded as well; as a consequence, the set of d-dimensional relations encompasses $13^d - 1$ elements. So, in a d-dimensional space, each dimension's relation from the set of Allen's relations $\mathcal{X}$ combines to form a comprehensive relation that captures the geometry of how hyperintervals H and H' are oriented concerning each other. This generalization enables us to qualitatively analyze the often complex spatial relationships present in larger-dimensional datasets, providing a

foundation for logical reasoning about the structure and patterns within the data.

Expanding upon the language and the semantics of interval temporal logic, we define now the logic $\mathcal{HS}^d$ as a generalization of $\mathcal{HS}$ to a d-dimensional space. Based on a set $\mathcal{P}$ of propositional variables, well-formed formulas in $\mathcal{HS}^d$ are constructed according to the syntax

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \langle X_1 \dots X_d \rangle \varphi,$$

where p represents a propositional variable from $\mathcal{P}$; as in the case of relations, $X_i \in \mathcal{X}$ for each dimension i. Operators of form $\langle X_1, \dots, X_d \rangle$ are called *existential*, indicating the possibility of transitioning to a different hyperinterval in the model. The remaining Boolean operators, as well as the *universal* version of each of the $13^d - 1$ existential operators are defined as shortcuts via duality (e.g., $[X_1, \dots, X_d]\varphi \equiv \neg\langle X_1, \dots, X_d \rangle \neg\varphi$).

To articulate the semantics of $\mathcal{HS}^d$, we define a *d-dimensional model* as an object of the type

$$M = \langle \mathbb{I}(\mathbb{D}^d), \phi \rangle$$

where $\mathbb{I}(\mathbb{D}^d)$ is the set of all hyperintervals across the d-dimensional space $\mathbb{D}^d$ and $\phi : \mathcal{P} \to 2^{\mathbb{I}(\mathbb{D}^d)}$ is a *valuation function* that maps each propositional variable to the set of hyperintervals where it holds. The truth of an $\mathcal{HS}^d$ formula φ at an hyperinterval H in model M follows an inductively defined set of conditions that align with typical Boolean operations and extend them with the existential constructs specific to $\mathcal{HS}^d$:

$$\begin{array}{lll} M, H \Vdash p & \text{iff} & H \in \phi(p), \text{ for each } p \in \mathcal{P}; \\ M, H \Vdash \neg\psi & \text{iff} & M, H \nVdash \psi; \\ M, H \Vdash \psi_1 \wedge \psi_2 & \text{iff} & M, H \Vdash \psi_1 \text{ and } M, H \Vdash \psi_2; \\ M, H \Vdash \langle X_1 \dots X_d \rangle \psi & \text{iff} & \exists H' \text{ s.t. } H R_{X_1 \dots X_d} H' \text{ and } M, H' \Vdash \psi. \end{array}$$

Observe that the *universal* operator $[U]^d$, that ensures a formula on every hyperinterval of a modal, is always definable; when the value d is clear from the context, we shall simply use $[U]$, instead.

When $d = 1$, $\mathcal{HS}^d$ becomes Halpern and Shoham's modal logic of time intervals $\mathcal{HS}$ [44]; its operators are those in Table 1, and formulas are interpreted over linear models (e.g., $\langle A \rangle p$ would be interpreted as *in some interval starting at the end of the current one, p holds*). When $d = 2$, $\mathcal{HS}^d$ becomes a modal logic of rectangles on a finite plane; its operators are of the type $\langle X_1, X_2 \rangle$, where both $\langle X_1 \rangle$ and $\langle X_2 \rangle$ are taken from Table 1, and contains, as definable sub-languages, well-known languages such as $\mathcal{RCC}8$ or $\mathcal{RCC}5$ (which are *logics of topological relations* [47]), with operators such as *partially overlapping*, $\langle PO \rangle$, or *externally connected*, $\langle EC \rangle$, among others. Versions of $\mathcal{HS}^d$ with more than two dimensions may not have previously studied from a computational point of view; nevertheless their usefulness in certain situations is undeniable.

With these logical constructs laid out, we observe that a d-dimensional dataset, described by a set of variables $\mathcal{V}$, can be viewed through the lens of a d-dimensional model when underpinned by a chosen set of feature extraction functions $\mathcal{F}$. The vocabulary of propositional variables is thus derived from combinations of these functions, variables, comparison operators, and real values, and defined as

$$\mathcal{P} = \{ f(V) \bowtie v \mid f \in \mathcal{F}, V \in \mathcal{V}, \bowtie \in \{<, \leq, =\}, v \in \mathbb{R} \}.$$

For each hyperinterval H within a specific instance I, these propositional variables evaluate a hypermatrix $I(V, H)$ composed of elements from the real numbers space across each dimension's interval range of variable V in hyperinterval H:

$$\phi(f(V) \bowtie v) = \{ H \mid H \in \mathbb{I}(\mathbb{D}^d) \text{ and } f(I(V, H)) \bowtie v \}.$$

When modal decision trees and modal random forests are utilized to probe d-dimensional datasets, they can unearth patterns expressible in $\mathcal{HS}^d$ [29–31,48]. Examples of such patterns, translated to natural language, range over all possible qualitative expressions concerning values of features on hyperintervals. Thus, in the temporal case, we may be looking at patterns/sentences such as

when the number of temperature peaks in a patient is greater than 3 in an interval of time, his/her blood pressure decreases quicker than 10mmHG/hour in some interval contained in the reference one,

in a hypothetical case in which patients under observation are described by a multivariate time series that includes *temperature* (*temp*) and *blood pressure* (*blood*) as variables, which would correspond to a formula such as

$$[U](peaks(temp) > 3 \rightarrow \langle D \rangle (deriv(blood) > 10)).$$

In the spatial case, examples as

when the maximum value of the frequency 'red' in an image if greater that 125 in some rectangle, there exist some adjacent rectangle with a minimum value of 'green' that is less than 64,

in a different hypothetical case, where instances are satellite images whose pixels are described by the amount of visible colors (*red* and *green*), would give rise to a formula of the type

$$[U](max(red) > 125 \rightarrow \langle EC \rangle (min(green) < 64)).$$

Modal formulas are discovered from a dataset in the same way in which propositional ones are: by systematic model checking of all instances. Such a computational process is expensive by nature, and this is why applicable learning methods are inexact by design (e.g., are greedy algorithms), and the number of variables, feature extraction functions, and hyperintervals to be examined has a great influence on the performances.

As we consider the scope of typical high-dimensional datasets, teeming with hundreds or even thousands of variables across multiple dimensions and subject to numerous candidate feature extraction functions, the resulting set of potential features spirals into the tens of thousands. Hence, a pivotal aspect of data processing is the thoughtful selection of a substantially smaller and more tractable subset of these features for in-depth analysis. This selection is not random but informed by the logic-driven examination of hyperintervals and their underlying relations in data instances within the multi-dimensional analytical space.

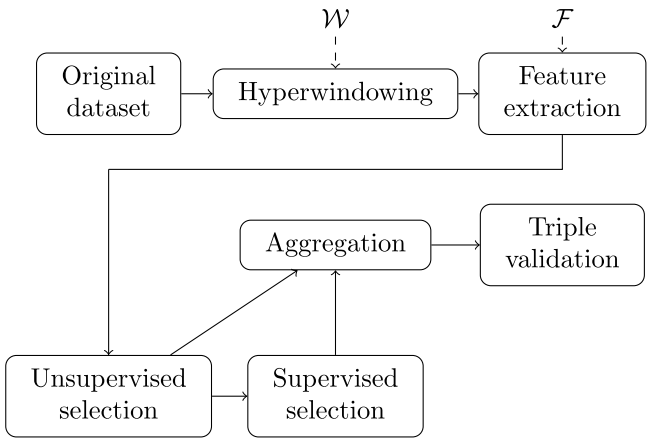


Fig. 1. A pictorial representation of dimensional feature extraction and selection.

Examples of modal symbolic learning. In the recent literature, modal symbolic learning has proven useful in several cases of extraction of information from dimensional data.

In [30], a dataset of cough and breath sample recordings of volunteer subjects, labeled with their COVID-19 status, was considered, and the problem of their automated classification using interval temporal decision trees and forests was studied; the obtained models not only outperformed the state-of-the-art obtained with the same dataset, but

their results were also superior to those of most non-symbolic techniques applied on other datasets. In [29], the problem under analysis was that of evaluating the EEG signal of subjects during the observation of artwork, and to extracting explicit rules that underly the electric signals at specific frequencies of specific electrodes, using temporal decision trees and forests (that is, modal decision trees and forest on $d = 1$ dimensional data); the resulting models showed performances comparable with the existing literature while being able to provide useful suggestions towards data understanding, including the importance of every electrode, the relevance of the exact frequency at which the information is carried, and the most informative measures (feature functions) that should be applied to the signal for such information to emerge. In [48] the vibration levels from gas turbines were considered with the aim of predicting episodes of turbine failure; again exploiting modal decision trees, it has been possible to select the most informative variables and produce relatively reliable models. Finally, in [49] the problem was to extract symbolic machine learning models from satellite data, that is, $d = 2$ dimensional data, using, whenever possible, spatial relations to enhance classification; modal trees and forests proved once more to be superior in terms of both statistical performances and interpretability of the results.

3. Dimensional feature extraction and selection

A systematic method for feature extraction and selection for dimensional data can be pictorially represented as in Fig. 1 and outlined as follows.

Hyperintervals. The number of different hyperintervals in a d-dimensional dataset in which each instance is delineated by N points across each axis is

$$(N(N+1)/2)^d.$$

Within these hyperintervals, evaluating variables under distinct functions intensifies the computational load. Directly addressing every potential combination of variables and functions over all hyperintervals is computationally impractical, if not outright infeasible. As a practicable alternative, we introduce a *hyperwindowing* approach. This method partitions a finite d-dimensional space into a manageable collection $\mathcal{W}$ of hyperintervals designated as *hyperwindows*. This set must cover every point within the space, ensuring comprehensive coverage without exhaustive enumeration, but they need not to be mutually exclusive. The selection of hyperwindow parameters determines the granularity and coverage of the dataset's features. By evaluating features within these hyperwindows, one can approximate the behaviors of the features across the broader sets of all hyperintervals with a controlled computational effort.

Upon establishing the hyperwindowing of the dataset, feature extraction begins. Extracting features across all hyperwindows results in a transformed dataset with dimensions influenced by the number of feature extraction functions applied. Normalization of extracted features is essential for maintaining comparability across different scales. This process results in a tabular dataset, offering a structured compilation of features encapsulating the intrinsic variations within hyperwindows. The tabular dataset—comprised of the application of each feature extraction function within each hyperwindow—serves as a foundation for subsequent selection steps.

An unsupervised feature selection step follows, leveraging normalization to ascertain an individual score for each variable-hyperwindow-function combination. A predefined fraction of the highest-scoring combinations is retained, condensing the dataset further. This step operates independently and, through its application, 'prunes' the dataset, eliminating features least likely to carry pertinent information. This dimensionality reduction facilitates a more focused investigation in the later stages of analysis, particularly as supervised methods tend to be computationally more intensive.

In the following phase, supervised feature selection applies a similar scoring and pruning process but within a target-oriented context. By considering the relationship between features and outcomes, supervised selection assigns relevance scores and filters the dataset to isolate the most predictive feature combinations. Observe that supervised selection can be skipped, for example when labels are not available or when they are ignored for some reasons.

Aggregation functions. Applying dimensional feature extraction and selection returns a set of triples of the type *(variable, hyperwindow, function)*. These, in turn, can be combined with *aggregation functions*. A very simple aggregation functions consists of evaluating a specific variable with the number of times in which it has been selected in the selected triples; the same can be done with windows and feature extraction function. Other, more elaborate aggregations can be designed, that take into account the specific scores. In general, aggregations may meld scores across multiple dimensions, allowing us to concentrate on the most influential variables, functions, or windows. As a final consideration, we observe that in the supervised case there exist popular feature selection methods based on hypothesis testing. Such a strategy in presence of several thousands of potential selectees, as in our cases, is cautioned against, given the risk of cumulative probabilistic errors leading to unreliable outcomes. However, the selected triples *can* be tested for their ability to distinguish classes (which is different from using the result of the test as part of the selection process). In the proposed framework, we include a *a posteriori* set of tests on a random sub-selection of triples, that allows us to evaluate the quality of the results.

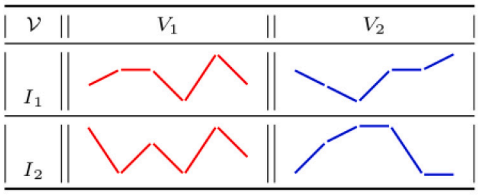
An example of this process in the temporal case can be seen in Fig. 2.

The outlined methodology balances computational feasibility and exhaustive feature inspection within d-dimensional datasets. Hyperwindowing represents an innovative compromise, reducing an otherwise insupportable analytical load into a tractable strategy. However, the size and arrangement of hyperwindows are pivotal and may necessitate empirical tuning to adequately capture the nuances of the dataset. Feature extraction and the dual-layered selection process exemplify a layered approach, focusing computational resources effectively from unspecific (unsupervised) to targeted (supervised) reduction. However, the appropriateness of chosen feature extraction functions remains crucial and context-dependent. Critically, while this approach enhances manageability, it may also introduce biases based on selecting hyperwindows and feature extraction functions. Furthermore, while helping with dimensionality, aggregation could conceal features with localized predictive power.

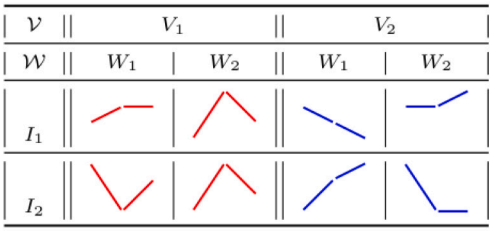
Implementation. We have developed and integrated the methodology discussed above into *Sole.jl* [43], an open-source framework dedicated to symbolic methods for data analysis, learning, and model analysis. In addressing the unsupervised step, we have adopted a standard variance-based filtering approach (*Var*), wrapping the well-known *Variance Threshold method* [50]. The framework offers *Mutual Information* (*MI*) [51] and *Fisher Score* (*FS*) [52] methods for the supervised case, which evaluate features according to their relationship with the class. Utilizing this technique involves a few critical steps for each application:

1. selecting the appropriate set of functions $\mathcal{F}$,
2. determining the percentage of features to retain after both unsupervised and supervised selection stages,
3. choosing how to aggregate the results for analysis, and
4. deciding the number of randomly chosen features, among the survived ones, to undergo a mean-based hypothesis testing step, to be used as validation phase.

This structured approach ensures that users can tailor the analysis process to their specific needs, promoting flexibility and efficiency in the application of symbolic learning-inspired analysis techniques.



(a) Two (raw) input multivariate time series.



(b) Hyperwindowing the time series in $\mathcal{W} = \{W_1, W_2\}$.

$\mathcal{V}$	V_1						V_2					
$\mathcal{W}$	W_1			W_2			W_1			W_2		
$\mathcal{F}$	f_1	f_2	f_3	f_1	f_2	f_3	f_1	f_2	f_3	f_1	f_2	f_3
I_1	$v^{(1)}_{111}$	$v^{(1)}_{112}$	$v^{(1)}_{113}$	$v^{(1)}_{121}$	$v^{(1)}_{122}$	$v^{(1)}_{123}$	$v^{(1)}_{211}$	$v^{(1)}_{212}$	$v^{(1)}_{213}$	$v^{(1)}_{221}$	$v^{(1)}_{222}$	$v^{(1)}_{223}$
I_2	$v^{(2)}_{111}$	$v^{(2)}_{112}$	$v^{(2)}_{113}$	$v^{(2)}_{121}$	$v^{(2)}_{122}$	$v^{(2)}_{123}$	$v^{(2)}_{211}$	$v^{(2)}_{212}$	$v^{(2)}_{213}$	$v^{(2)}_{221}$	$v^{(2)}_{222}$	$v^{(2)}_{223}$

(c) Feature extraction using $\mathcal{F} = \{f_1, f_2, f_3\}$.

$\mathcal{V}$	V_1						V_2					
$\mathcal{W}$	W_1			W_2			W_1			W_2		
$\mathcal{F}$	f_1	f_2	f_3	f_1	f_2	f_3	f_1	f_2	f_3	f_1	f_2	f_3
I_1	$v^{(1)}_{111}$	$v^{(1)}_{112}$	$v^{(1)}_{113}$	$v^{(1)}_{121}$	$v^{(1)}_{122}$	$v^{(1)}_{123}$	$v^{(1)}_{211}$	$v^{(1)}_{212}$	$v^{(1)}_{213}$	$v^{(1)}_{221}$	$v^{(1)}_{222}$	$v^{(1)}_{223}$
I_2	$v^{(2)}_{111}$	$v^{(2)}_{112}$	$v^{(2)}_{113}$	$v^{(2)}_{121}$	$v^{(2)}_{122}$	$v^{(2)}_{123}$	$v^{(2)}_{211}$	$v^{(2)}_{212}$	$v^{(2)}_{213}$	$v^{(2)}_{221}$	$v^{(2)}_{222}$	$v^{(2)}_{223}$

(d) Unsupervised and supervised feature selection.

Fig. 2. Layered pipeline for feature extraction and feature selection for high-dimensional datasets.

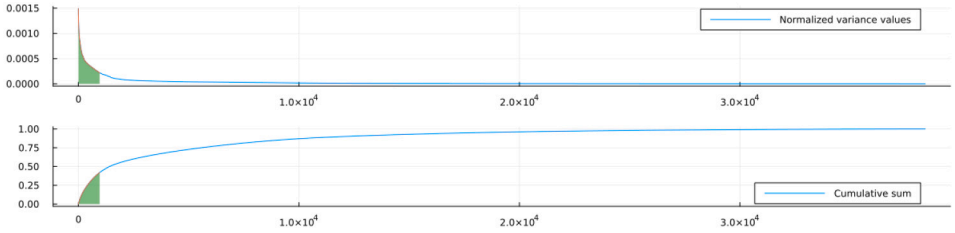
4. Results and discussion

To demonstrate the efficacy and effectiveness of our methodology and its implementation, we consider two distinct dimensional cases: a 1-dimensional (temporal data) and a 2-dimensional (spatial data) scenario.

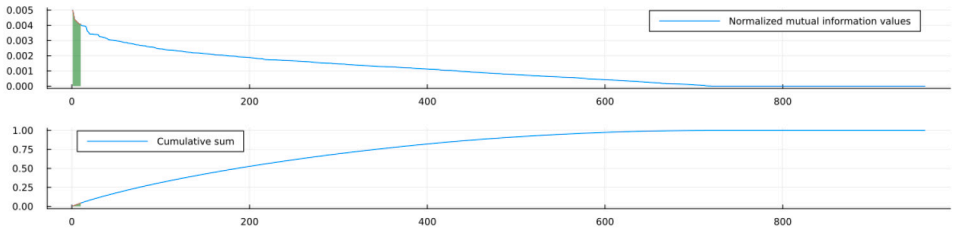
Temporal case. Focusing on the cade of temporal data, our analysis encompasses a dataset of *electroencephalographic* (*EEG*) recordings derived from an experiment investigating the corollary discharge process within the nervous system. This process is pivotal for differentiating between stimuli originating internally versus those from external sources. The experiment engaged subjects both with and without schizophrenia in tasks designed to elicit this distinction, by having them either press a button to generate a tone, passively listen to a tone, or press a button without tone generation. These tasks resulted in distinct datasets: T_1, T_2, and T_3, each reflecting a unique order of actions. Schizophrenia's link to potential disruptions in this discharge process underpins our expectation of observable differences in EEG signals between control subjects and those with schizophrenia. This hypothesis is supported by prior research indicating a suppression of the negative deflection in EEG brainwaves by control subjects upon self-generated tone, an effect not mirrored in individuals with schizophrenia [53].

The data, available from [54,55], encompasses 32 control subjects and 49 patients with schizophrenia, for a total of 81 subjects. It is part of a bigger dataset already used in a broader study that included fMRI data along with EEG [56]. Each participant, on average, partook in approximately 285 trials, resulting in a relatively rich dataset in terms instances (trials), which are labeled across two classes ($k = 2$). Previous works used the same dataset (e.g., see [57,58]). The pre-processing regimen applied to this dataset included filtering, epoching, baseline correction, canonical correlation analysis, and outlier rejection. Each EEG recording, capturing data from 64 electrodes (channels) positioned according to the *10–20 system* and recorded at a 1024Hz sampling rate over a period of 3 s, provided a comprehensive view of brain activity through 3072 data points per recording.

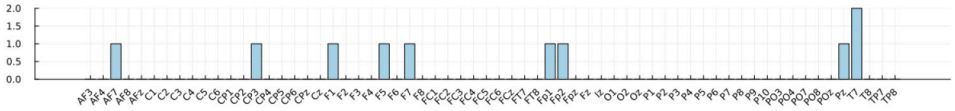
The application of the Short-Time Fourier Transform (STFT) on each channel facilitated the extraction of 40 1 Hz-wide frequency bands from 0 to 40 Hz. This operation expanded our feature set to 2560 ($n = 64 \cdot 40$) per recording. Following this, the dataset was segmented according to the three distinct tasks described earlier, yielding subsets with 7836, 7631, and 7734 instances (m values), respectively. This segmentation



(a) Unsupervised feature selection.



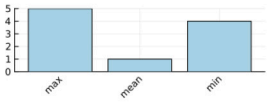
(b) Supervised feature selection.



(c) Electrode occurrences.

(d) Frequency band occurrences.

(e) Window occurrences.



(f) Measure occurrences.

$p \leq 0.05$
Pz F1.5 W2 max (4.0×10^{-3})
Fp1 F1.5 W2 min (1.9×10^{-3})
T7 F0.5 W4 min (3.9×10^{-3})
AF7 F1.5 W2 max (1.8×10^{-3})
F7 F0.5 W5 max (2.2×10^{-5})
CP3 F1.5 W1 min (1.3×10^{-3})
T7 F1.5 W1 max (2.4×10^{-3})

(g) Feature validation.

Fig. 3. Results for the T_1 dataset.

reflects the repeated execution of each task by the subjects. The size of each task-specific dataset approximates 13 GB in a CSV format.

To capture the temporal dynamics within each trial, we defined five distinct time windows,

$$\mathcal{W} = \{[1, 639], [610, 1247], [1217, 1855], [1825, 2462], [2433, 3072]\},$$

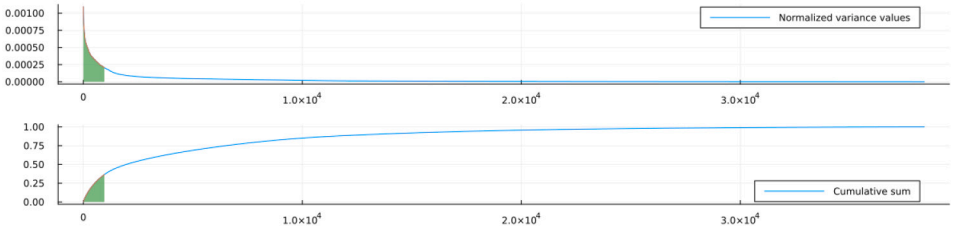
($|\mathcal{W}| = 5$). This structuring ensures that all trials are aligned at their onset, with any significant event occurring 150 ms post-initiation. Furthermore, we considered three statistical measures—mean, maximum, and minimum ($\mathcal{F} = \{mean, maximum, minimum\}$, with $l = 3$)—to enrich our feature analysis, pushing the potential feature count to 38400 per dataset.

The feature selection journey commenced with an unsupervised step, employing variance analysis (*Var*) to narrow down our expansive feature set to 2.5% of its original size, resulting in 960 features. This was further refined through a supervised phase, leveraging Mutual Information (*Mutual Information*) to pinpoint the top 10 features, representing 1% of the previously selected subset. These final features underwent a rigorous validation process using the Mann–Whitney U-test [59], with a significance threshold set at $\alpha = 0.05$, ensuring their statistical relevance and robustness for our study's objectives.

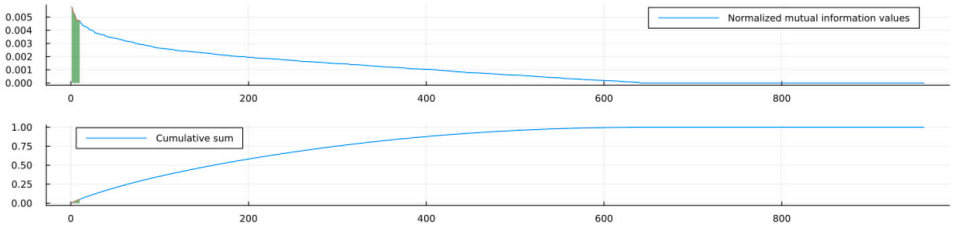
Our analytical results are detailed in Figs. 3, 4, and 5. Figs. 3(a), 4(a), and 5(a) demonstrate a significant drop in variance score moving from left to right across the figures, highlighting a rapid decline in feature relevance. Specifically, the initial unsupervised selection phase accounted for 42%, 37%, and 35% of the total variance for T_1, T_2, and T_3, respectively. A deeper analysis revealed that achieving 50% variance explanation would necessitate a much larger feature set (1410 features for T_1, 1957 for T_2, and 2129 for T_3, respectively), validating our selection of 2.5% post-unsupervised step as both efficient and adequate. An analogous examination reveals that, as depicted in Figs. 3(b), 4(b), and 5(b), the decline in normalized mutual information scores is more gradual compared to the descent observed in variance scores. Incorporating this understanding, it becomes evident that the mutual information scores for the features that persisted beyond the unsupervised step are computed by assessing their relationship with the target variable. This methodological approach underscores the gradual decline in scores, reflecting the retained features' relevance and their informativeness regarding the target.

The features have the form (V, W, f), where V, in turn, is a pair

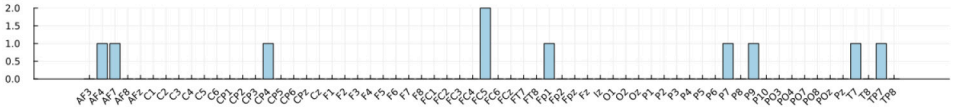
$(electrode, frequency)$.



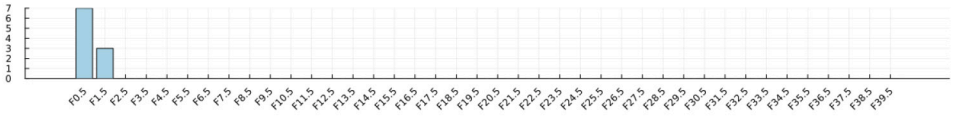
(a) Unsupervised feature selection.



(b) Supervised feature selection.

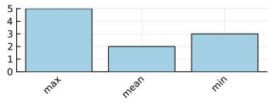


(c) Electrode occurrences.



(d) Frequency band occurrences.

(e) Window occurrences.



(f) Measure occurrences.

$p \leq 0.05$
CP4 F0.5 W1 mean (2.6×10^{-2})
AF4 F0.5 W5 max (7.7×10^{-4})
TP7 F0.5 W1 max (2.6×10^{-2})
FC5 F0.5 W4 max (1.4×10^{-3})

(g) Feature validation.

Fig. 4. Results for the T_2 dataset.

Thus, the final set of 10 features can be aggregated in four different ways, both by occurrences and score (after the supervised filter). This approach leads to two different groups of four diagrams each, respectively in Figs. 3(c), 3(d), 3(e), 3(f), 4(c), 4(d), 4(e), 4(f), 5(c), 5(d), 5(e), and 5(f). As it can be seen, the electrodes that occurred the highest number of times in the 10 selected features are

$$AF7, CP3, F1, F5, F7, Fp1, Fp2, Pz, T7$$

for T_1,

$$AF4, AF7, CP4, FC5, Fp1, P7, P9, T7, TP7$$

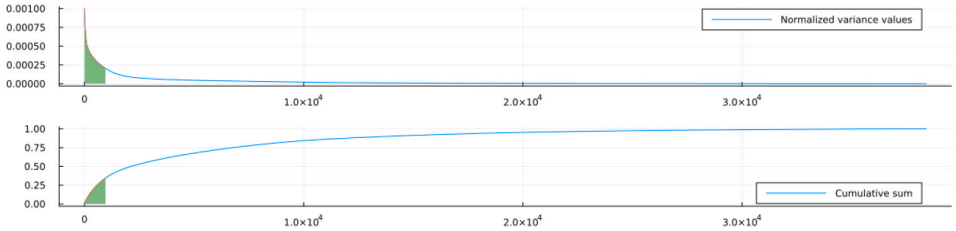
for T_2, and

$$AF7, Cz, F4, F6, F7, FC5, FC6, FT7, Fp2, Fpz$$

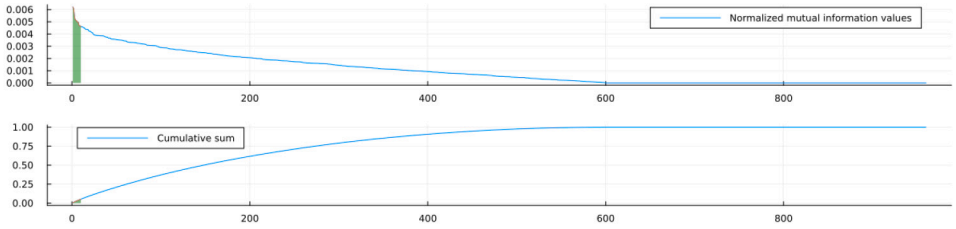
for T_3; such electrodes showed the most amount of information on the frequency bands centered on 0.5 and 1.5 Hz. Furthermore, the temporal windows preceding the event onset emerged as particularly informative, underscoring the predictive potential of pre-event brain activity. Finally, the 10 selected features underwent a Mann–Whitney U-test. Figs. 3(g), 4(g), and 5(g), shows that 7, 4, and 3 out of 10 features have passed the test for T_1, T_2, and T_3, respectively. It should be noted that the parameter choices were made in advance; however, alternative selections were possible. The primary aim of this experiment is to demonstrate the viability of the protocol.

This meticulous approach to feature selection and validation through statistical testing underscores the robustness of our methodology. This work could represent the initial part of another research. Indeed, by focusing on statistically significant EEG patterns and their correlation with specific neural processes, our study not only sheds light on the neural underpinnings of schizophrenia but also advances our understanding of the brain's ability to differentiate between self-generated and external stimuli.

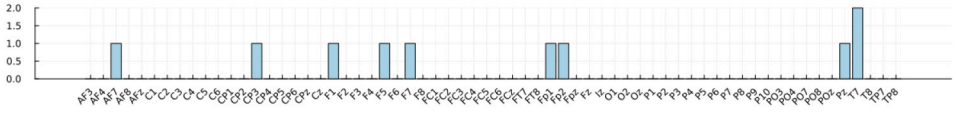
Spatial case. To prove the effectiveness of our approach in spatial case as well, we chose three multi-class datasets: *Indian Pines (IP)*, *Pavia University (PU)*, and *Pavia Centre (PC)*, each a benchmark in the realm of *land cover classification*. This task entails assigning each pixel of remotely sensed imagery a label indicative of land usage, such as forest, urban area, or crop field. Although resembling standard image segmentation, land cover classification is primarily treated as an image classification challenge, where the imagery, typically sourced from satellite or aerial sensors, is hyperspectral. As such, this means it comprises a broad spectrum of spatial variables (*spectral channels*), each representing signal strength of a specific electromagnetic wavelength, with the classification of a single pixel relying on its immediate



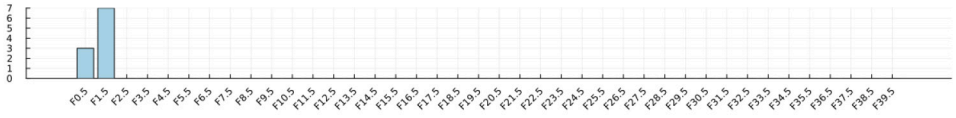
(a) Unsupervised feature selection.



(b) Supervised feature selection.

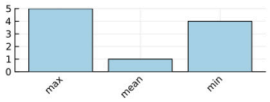


(c) Electrode occurrences.



(d) Frequency band occurrences.

(e) Window occurrences.



(f) Measure occurrences.

$p \leq 0.05$
F7 F0.5 W4 mean (2.4×10^{-2})
AF7 F1.5 W1 min (5.5×10^{-4})
FT7 F0.5 W4 min (2.5×10^{-2})

(g) Feature validation.

Fig. 5. Results for the T_3 dataset.

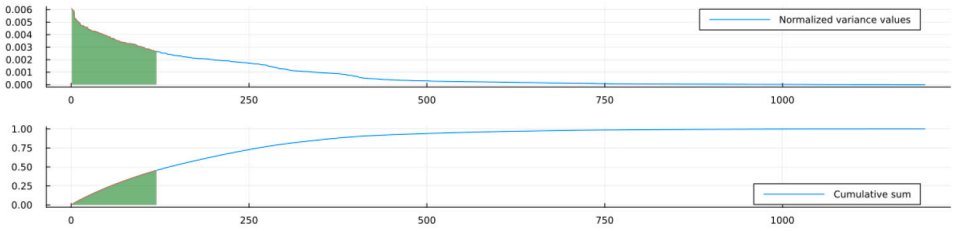
neighborhood's pixel ensemble. Note that, similarly to the case of EEG frequencies, spectral channels also delineate electromagnetic bands.

Diving into the specifics, the IP dataset comprises 1600 images ($m = 1600$), each structured into a 5×5 pixel matrix, with 200 spectral variables $C_1, \ldots, C_{200}$ spanning from 400 to 2500 nanometers ($n = 200$), organized into 16 distinct classes ($k = 16$). For the PU dataset, we encounter 103 spectral variables ($C_1, \ldots, C_{103}$) ranging from 430 to 860 nanometers ($n = 103$), spread across 900 images ($m = 900$) and segmented into 9 classes ($k = 9$). The PC dataset mirrors this structure with 102 variables ($C_1, \ldots, C_{102}$), from 430 to 860 nanometers ($n = 102$), also divided among 9 classes ($k = 9$) across 900 images ($m = 900$). The datasets occupy 61, 18, and 18 megabytes, respectively. In our analysis, a 3×3 window $W_1 = ([2, 2], [4, 4])$ centered around the pixel, as well as the 5×5 window $W_2 = ([1, 1], [5, 5])$ enclosing the whole neighborhood are used ($\mathcal{W} = \{W_1, W_2\}$), alongside three extraction functions ($\mathcal{F} = \{mean, maximum, minimum\}$, with $l = 3$), thus yielding a tailored feature space for each dataset—1200 for IP, 618 for PU, and 612 for PC, respectively.

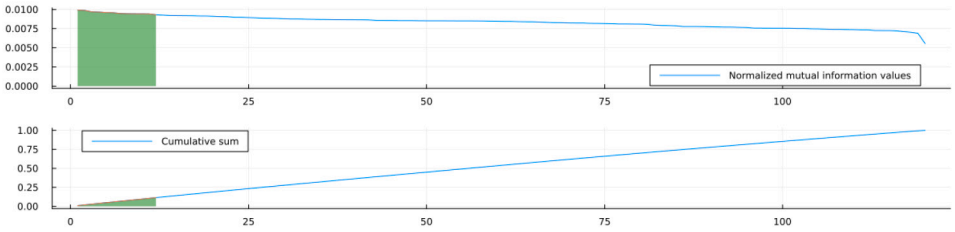
Venturing into the analytical depths, outlined in Figs. 6, 7, and 8, the variance reduction trajectory in the spatial datasets exhibited a more gradual decline compared to the temporal scenario. Post-unsupervised filtering via variance led to a potential feature pool—120 for IP and 62 for both PU and PC. A further refined examination through a supervised Mutual Information step distilled this to a core set of 12 features per dataset—10% for IP and 20% for both PU and PC of each feature pool (rounded to the lower integer).

Notably, the initial (unsupervised) reduction explained 46%, 24%, and 27% of the total variance for IP, PU, and PC, respectively. A closer look revealed that a fraction of these—138 for IP, 176 for PU, and 137 for PC—would suffice to account for 50% of the variance, suggesting a strategic, minimalistic approach to feature selection could be highly effective.

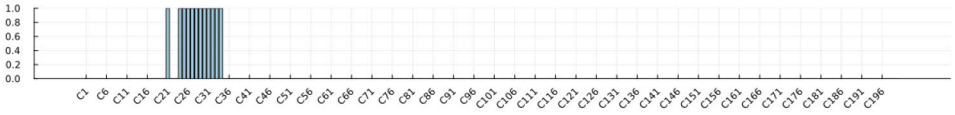
The feature analysis shed light on distinctive spectral bands of significance across the datasets. IP's analysis delineated a single critical band corresponding to channels C_{21}–C_{36}, which lie at the infrared verge (wavelengths between 596–738 nanometers). The narrower spatial window (W_1) and the predominance of *maximum* function as feature extractors emerged as pivotal in discerning the land cover classes. In contrast, PU highlighted a significant band corresponding to channels C_{14}–C_{32}, which corresponds to the visible colors cyan and green (wavelengths between 488–563 nanometers), with classifications becoming more discernible within the wider spatial window (W_2) and under the lens of *maximum* and *mean* functions. The PC dataset echoed this pattern, pinpointing a singular impactful infrared band (C_{89}–C_{102}, with wavelengths between 802–856 nanometers), with class clarity benefiting from the wider spatial window (W_2) and similar function preferences. Our protocol selected 10 features at random from the 12 core feature set for validation. Mann–Whitney U-test validated the 10



(a) Unsupervised feature selection.



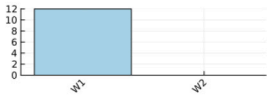
(b) Supervised feature selection.



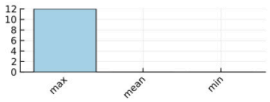
(c) Hyperspectral band occurrences.

$p \leq 0.01$
C32 W1 maximum (1.3×10^{-61})
C34 W1 maximum (6.8×10^{-59})
C25 W1 maximum (2.7×10^{-62})
C21 W1 maximum (9.0×10^{-63})
C31 W1 maximum (2.6×10^{-61})
C26 W1 maximum (2.7×10^{-62})
C30 W1 maximum (7.7×10^{-62})
C28 W1 maximum (3.4×10^{-62})
C33 W1 maximum (2.5×10^{-61})
C24 W1 maximum (2.0×10^{-62})

(f) Feature validation.



(d) Window occurrences.



(e) Measure occurrences.

Fig. 6. Results for the IP dataset.

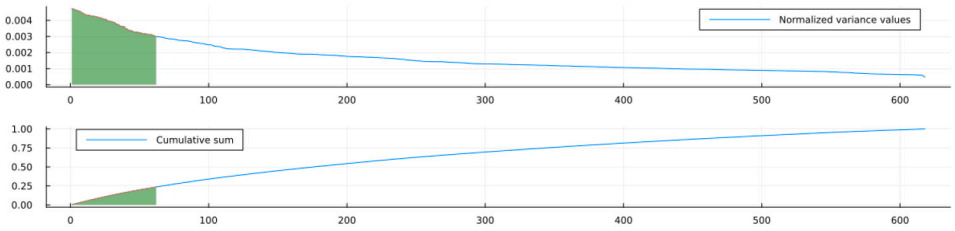
selected features—as in the temporal scenario. Figs. 6(f), 7(f), and 8(f) show that all 10 features have passed the test for all datasets, indicating the effectiveness of the selection protocol.

This intricate analysis underscores the complexity inherent in spatial data interpretation. It exemplifies the synergy between spectral bands, spatial windows, and statistical functions in unraveling the mosaic of land cover types. Remarkably, this research addresses a gap in the existing literature, which needs to be analyzed for image segmentation and classification features, even more so than in the temporal case. This gap highlights the potential for our work to mark the beginning of a new area of exploration, introducing novel approaches to the field. The unerring validation of the selected features accentuates the precision of our methodology, promising a refined lens through which the nuances of land cover can be discerned. Thereby, our findings not only advance the field of hyperspectral image analysis in land cover classification but also propose a pioneering direction for future research, underscoring its novelty and significance.

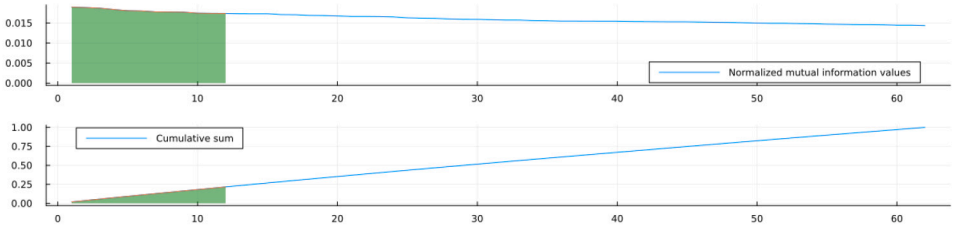
5. Conclusions

This research represents a significant stride forward in big data analytics, mainly through its novel incorporation of symbolic methodologies for high-dimensional data analysis. The symbolic perspective adopted here enriches our methodological arsenal and deepens our conceptual understanding of data's inherent structures and relationships. By developing and sharing an open-source software package, we have laid a robust platform for applying, examining, and extending these symbolic techniques within the broader scientific and technological community.

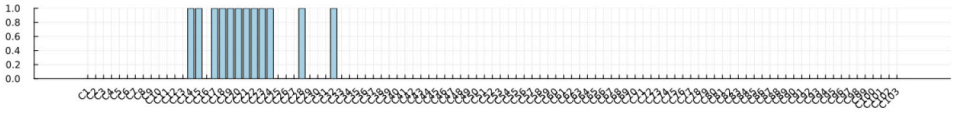
From a symbolic standpoint, our work tackles the multifaceted challenges posed by high-dimensional datasets, including issues related to scalability, noise, and feature relevance. This approach has facilitated the discovery of more nuanced, interpretable models with significant promise for complex data analysis across various domains. The symbolic dimension of our research invites a reevaluation of traditional data analysis paradigms, urging a shift towards methodologies that prioritize not only computational efficiency but also interpretability and theoretical soundness. While in our experiments we limited ourselves to apply specific univariate filters, both unsupervised and supervised, our framework is designed to incorporate essentially any technique of the same type, including, but not limited to, *Mean Absolute Difference* and *Dispersion ratio*, as well as the entire range of typical correlation indexes, including *Pearson's*, *Spearman's*, and *Kendall's*. In fact, the principle at the core of our approach is begin able to study dimensional data exactly as it is usually done in the tabular case.



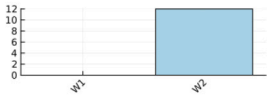
(a) Unsupervised feature selection.



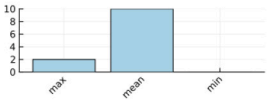
(b) Supervised feature selection.



(c) Hyperspectral band occurrences.



(d) Window occurrences.



(e) Measure occurrences.

$p \leq 0.01$
C22 W2 mean (7.3×10^{-60})
C19 W2 mean (7.2×10^{-60})
C14 W2 mean (7.1×10^{-60})
C28 W2 maximum (7.7×10^{-60})
C24 W2 mean (7.5×10^{-60})
C21 W2 mean (7.3×10^{-60})
C15 W2 mean (7.1×10^{-60})
C20 W2 mean (7.3×10^{-60})
C17 W2 mean (7.2×10^{-60})
C23 W2 mean (7.4×10^{-60})

(f) Feature validation.

Fig. 7. Results for the PU dataset.

Looking to the future, the methodologies introduced in this research set the stage for a new era of data analytics, where symbolic approaches play a central role in unraveling the complexities of big data. The open-source nature of our contribution encourages a collaborative approach to innovation, inviting researchers and practitioners to explore the full potential of symbolic data analysis in their work.

In sum, our research underscores the critical importance of symbolic methodologies in advancing the field of big data analytics. It represents a call to action for the community to embrace these approaches, fostering an environment where data's most profound insights can be uncovered. As we continue to explore this promising frontier, our collective efforts will undoubtedly lead to more sophisticated, insightful, and impacting data analysis techniques, shaping the future of how we understand and interact with the vast digital universe around us.

CRediT authorship contribution statement

P. Cavina: Writing – review & editing, Validation, Software. **F. Manzella:** Writing – review & editing, Software, Methodology, Formal analysis. **G. Pagliarini:** Software, Methodology, Investigation. **G. Sciavicco:** Writing – original draft, Validation, Methodology, Investigation, Conceptualization. **I.E. Stan:** Writing – review & editing, Supervision, Conceptualization.

Declaration of Generative AI Use

During the preparation of this work no generative AI tool was used, and the authors are fully responsible of the content.

Declaration of competing interest

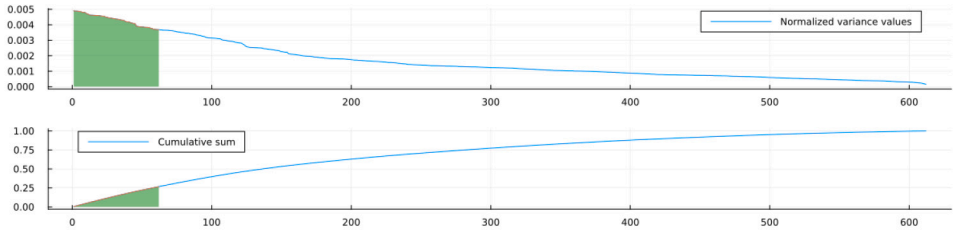
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

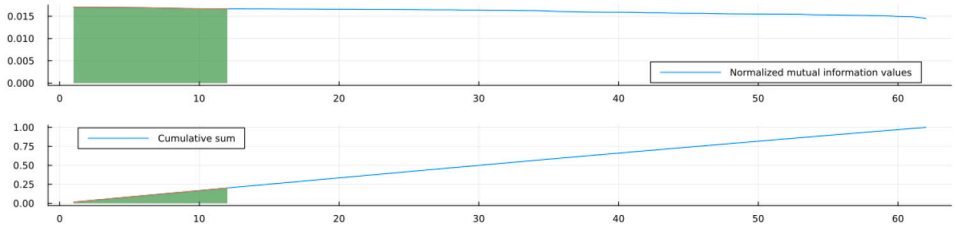
This research was partially funded by INdAM - GNCS Project *Semantic neuro-symbolic contract for the rigorous reachability and CPS certifications* (CUP: E53C25002010001); Guido Sciavicco and Ionel Eduard Stan are members of INdAM.

Data availability

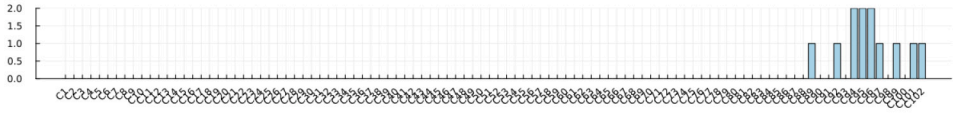
All data used in this paper is public and openly accessible.



(a) Unsupervised feature selection.



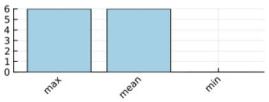
(b) Supervised feature selection.



(c) Hyperspectral band occurrences.

$p \leq 0.01$
C95 W2 mean (1.1×10^{-57})
C89 W2 mean (7.9×10^{-58})
C96 W2 maximum (5.3×10^{-58})
C94 W2 maximum (8.8×10^{-58})
C97 W2 mean (7.3×10^{-58})
C94 W2 mean (1.0×10^{-57})
C99 W2 maximum (4.2×10^{-58})
C92 W2 mean (9.1×10^{-58})
C101 W2 maximum (2.9×10^{-58})
C96 W2 mean (1.0×10^{-57})

(d) Window occurrences.



(e) Measure occurrences.

(f) Feature validation.

Fig. 8. Results for the PC dataset.

References

[1] Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. Int J Inf Manage 2015;35(2):137–44.

[2] Tanwar S, Ramani T, Tyagi S. Dimensionality reduction using PCA and SVD in big data: A comparative case study. In: Proc. of the first international conference on future internet technologies and trends. 2018, p. 116–25.

[3] Liu H, Motoda H. Feature selection for knowledge discovery and data mining. Kluwer Academic Publishers; 1998.

[4] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.

[5] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 2005;17(4):491–502.

[6] Alelyani S, Tang J, Liu H. Feature selection for clustering: A review. In: Data clustering: algorithms and applications. CRC Press; 2013, p. 29–60.

[7] Tang J, Alelyani S, Liu H. Feature selection for classification: A review.. In: Data classification: algorithms and applications. CRC Press; 2014, p. 37–64.

[8] Vergara JR, Estévez PA. A review of feature selection methods based on mutual information. Neural Comput Appl 2014;24(1):175–86.

[9] Ang JC, Mirzal A, Haron H, Hamed HNA. Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. IEEE/ACM Trans Comput Biol Bioinform 2016;13(5):971–89.

[10] Miao J, Niu L. A survey on feature selection. Procedia Comput Sci 2016;91:919–26.

[11] Sheikhpour R, Sarram MA, Gharaghani S, Chahooki MAZ. A survey on semi-supervised feature selection methods. Pattern Recognit 2017;64(C):141–58.

[12] Li J, Liu H. Challenges of feature selection for big data analytics. IEEE Intell Syst 2017;32(2):9–15.

[13] Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. Feature selection: A data perspective. ACM Comput Surv 2017;50(6).

[14] Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF. A review of unsupervised feature selection methods. Artif Intell Rev 2020;53(2):907–48.

[15] Abdulwahab H, Ajitha S, Saif M. Feature selection techniques in the context of big data: taxonomy and analysis. Appl Intell 2022;52(12):13568–613.

[16] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing 2018;300:70–9.

[17] Solorio-Fernández S, Carrasco-Ochoa J, Martínez-Trinidad J. A systematic evaluation of filter unsupervised feature selection methods. Expert Syst Appl 2020;162:1–26.

[18] Huang S. Supervised feature selection: A tutorial. Artif Intell Res 2015;4:22–37.

[19] Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng 2014;40(1):16–28.

[20] Rodríguez Diez J, González C, Boström H. Boosting interval based literals. Intell Data Anal 2001;5(3):245–62.

[21] Yamada Y, Suzuki E, Yokoi H, Takabayashi K. Decision-Tree Induction from Time-Series Data Based on a Standard-Example Split Test. In: Proce. of the 12th international conference on machine learning. 2003, p. 840–7.

[22] Brunello A, Marzano E, Montanari A, Sciavicco G. J48SS: a novel decision tree approach for the handling of sequential and time series data. Computers 2019;8(1):21.

[23] Jha S, Tiwari A, Seshia SA, Sahai T, Shankar N. TeLEx: learning signal temporal logic from positive examples using tightness metric. Form Methods Syst Des 2019;54(3):364–87.

[24] Li X, Claramunt C. A spatial entropy-based decision tree for classification of geographical information. Trans GIS 2006;10(3):451–67.

[25] Stan IE. foundations of modal symbolic learning (Ph.D. thesis), University of Parma; 2023, Available at URL https://www.repository.unipr.it/bitstream/1889/5219/5/main.pdf.

[26] Brunello A, Sciavicco G, Stan IE. Interval Temporal Logic Decision Tree Learning. In: Proceedings of the 16th European conference on logics in artificial intelligence (JELIA 2019). Lecture notes in computer science, vol. 11468, Springer; 2019, p. 778–93.
[27] Della Monica D, Pagliarini G, Sciavicco G, Stan I. Decision Trees with a Modal Flavor. In: Proc. of the 21st international conference of the Italian association for artificial intelligence (aIxIA). Lecture notes in computer science, vol. 13796, Springer; 2022, p. 47–59.
[28] Manzella F, Pagliarini G, Sciavicco G, Stan IE. Efficient modal decision trees. In: Proceedings of the 22nd international conference of the Italian association for artificial intelligence (aIxIA 2023). Lecture notes in computer science, vol. 14318, Springer; 2023, p. 381–95.
[29] Mazzacane S, Coccagna M, Manzella F, Pagliarini G, Sironi V, Gatti A, Caselli E, Sciavicco G. Towards an objective theory of subjective liking: A first step in understanding the sense of beauty. PloS One 2023;18:e0287513.
[30] Manzella F, Pagliarini G, Sciavicco G, Stan I. The voice of COVID-19: breath and cough recording classification with temporal decision trees and random forests. Artifcial Intell Med 2023;137:102486.
[31] Pagliarini G, Sciavicco G. Decision tree learning with spatial modal logics. In: Proc. of the 12th international symposium on games, automata, logics, and formal verification (gandALF). EPTCS, vol. 346, 2021, p. 273–90.
[32] Lubba C, Sethi S, Knaute P, Schultz S, Fulcher B, Jones N. Catch22: Canonical time-series characteristics - selected through highly comparative time-series analysis. Data Min Knowl Discov 2019;33(6):1821–52.
[33] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 1980;28(4):357–66.
[34] Dhal P, Azad C. A comprehensive survey on feature selection in the various fields of machine learning. Appl Intell 2022;52(4):4543–81.
[35] Crone SF, Kourentzes N. Feature selection for time series prediction–a combined filter and wrapper approach for neural networks. Neurocomputing 2010;73(10–12):1923–36.
[36] Sun Y, Li J, Liu J, Chow C, Sun B, Wang R. Using causal discovery for feature selection in multivariate numerical time series. Mach Learn 2015;101(1):377–95.
[37] Kathirgamanathan B, Cunningham P. A feature selection method for multi-dimension time-series data. In: International workshop on advanced analytics and learning on temporal data. Springer; 2020, p. 220–31.
[38] Tsimpiris A, Kugiumtzis D. Feature selection for classification of oscillating time series. Expert Syst 2012;29(5):456–77.
[39] Bolon-Canedo V, Remeseiro B. Feature selection in image analysis: a survey. Artif Intell Rev 2020;53(4):2905–31.
[40] Raj RJS, Shobana SJ, Pustokhina IV, Pustokhin DA, Gupta D, Shankar K. Optimal feature selection-based medical image classification using deep learning model in internet of medical things. IEEE Access 2020;8:58006–17.
[41] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825–30.
[42] Kuhn M, Johnson K. Applied predictive modeling. Springer; 2013.
[43] Manzella F, Pagliarini G, Sciavicco G, Stan IE. Sole.jl – symbolic learning in julia. 2013, https://github.com/aclai-lab/Sole.jl.
[44] Halpern J, Shoham Y. A Propositional Modal Logic of Time Intervals. J ACM 1991;38(4):935–62.
[45] Balbiani P, Condotta J, Fariñas del Cerro L. A model for reasoning about bidimensional temporal relations. In: Proc. of the 6th international conference on principles of knowledge representation and reasoning (kR'98). Morgan Kaufmann; 1998, p. 124–30.
[46] Allen JF. Maintaining Knowledge about Temporal Intervals. Commun ACM 1983;26(11):832–43.
[47] Lutz C, Wolter F. Modal logics of topological relations. Log Methods Comput Sci 2006;2. URL https://api.semanticscholar.org/CorpusID:1765850.
[48] Bechini G, Losi E, Manservigi L, Pagliarini G, Sciavicco G, Stan IE, Venturini M. Statistical rule extraction for gas turbine trip prediction. J Eng Gas Turbines Power 2023;145(5).
[49] Pagliarini G, Sciavicco G. Interpretable land cover classification with modal decision trees. Eur J Remote Sens 2023;56(1):1–25.
[50] Pedregosa F, et al. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825–30.
[51] Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Phys Rev E 2004;69(6):066138.
[52] Duda RO, Hart PE, Stork DG. Pattern classification. second ed.. New York: Wiley; 2001.
[53] Näätänen R, Picton T. The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. Psychophysiology 1987;24(4):375–425.
[54] Broach KDC. EEG data from basic sensory task in schizophrenia. 2024, URL https://www.kaggle.com/datasets/broach/button-tone-sz. [Accessed 15 September 2024].
[55] Broach KDC. EEG data from sensory task in schizophrenia (2). 2024, URL https://www.kaggle.com/datasets/broach/buttontonesz2. [Accessed 15 September 2024].
[56] National Institute of Mental Health. Predictive coding abnormalities in psychosis: EEG and fMRI. 2016, URL https://reporter.nih.gov/project-details/9187052.
[57] Barros C, Roach B, Ford JM, Pinheiro AP, Silva CA. From sound perception to automatic detection of schizophrenia: an EEG-based deep learning approach. Front Psychiatry 2022;12:813460.
[58] Siuly S, Khare SK, Bajaj V, Wang H, Zhang Y. A computerized method for automatic detection of schizophrenia using EEG signals. IEEE Trans Neural Syst Rehabil Eng 2020;28(11):2390–400.
[59] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947;50–60.