A compact information-theoretic framework for texture classification: Hilbert curves, amplitude-aware permutation entropy, and explainability

Cite as: Chaos 36, 083102 (2026); doi: 10.1063/5.0341167
Submitted: 29 April 2026 · Accepted: 13 July 2026 ·
Published Online: 3 August 2026

View Online Export Citation CrossMark

Ma. Belén Arouxet,[1,a] Aurelio F. Bariviera,[2,b] Roberta Hansen,[3,c] and Verónica E. Pastor[3,4,d]

AFFILIATIONS

[1] Facultad de Ciencias Exactas, Centro de Matemática de La Plata, Universidad Nacional de La Plata, B1900 La Plata, Argentina
[2] Departament de Gestió d'Empreses, ECO-SOS, Universitat Rovira i Virgili, Av. Universitat 1, 43204 Reus, Spain
[3] Facultad de Ingeniería, Departamento de Matemáticas, Universidad de Buenos Aires, C1063ACV Buenos Aires, Argentina
[4] Facultad de Ingeniería, Departamento de Ciencias Básicas, Universidad Nacional de La Plata, B1900 La Plata, Argentina

Note: This paper is part of the Special Topic dedicated to the theory and applications of ordinal patterns.

[a] belen@mate.unlp.edu.ar
[b] **Author to whom correspondence should be addressed:** aurelio.fernandez@urv.cat
[c] rhansen@fi.uba.ar
[d] vpastor@fi.uba.ar

ABSTRACT

Texture classification requires characterizing spatial arrangements of pixel intensities, such as periodicity, directionality, and surface roughness. This task is critical across diverse domains, including remote sensing, materials science, and biomedical imaging. In this paper, we propose a two-step information-theoretic framework for texture discrimination. First, each image is transformed into a one-dimensional time series using a space-filling Hilbert curve. This approach preserves spatial locality while avoiding the introduction of a privileged scanning direction. Second, we extract a comprehensive set of eight complementary quantifiers from the resulting series: permutation entropy, statistical complexity, Fisher information (under both lexicographic and colexicographic orderings), Wasserstein distances between the ordinal-pattern distribution and the uniform distribution, weighted permutation entropy, and amplitude-aware permutation entropy. These scalar features are then used to train a support vector machine with a radial basis function kernel, utilizing nested cross-validation for hyperparameter optimization. We validate this framework on the Kylberg texture database, a benchmark consisting of 280 images across 28 classes. Our results demonstrate that while the classical complexity-entropy causality plane offers some discriminative power, amplitude-sensitive quantifiers (specifically weighted permutation entropy) are the primary drivers of performance. These findings underscore the necessity of encoding amplitude information alongside ordinal patterns for effective texture characterization.

© 2026 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1063/5.0341167

Texture classification is a fundamental problem in image analysis. This approach discards amplitude information, which limits its discriminative power. This work advances the field by proposing a compact, interpretable, and statistically rigorous framework that integrates Hilbert curve scanning with an extended set of eight complementary quantifiers, including amplitude-sensitive measures, Fisher information, and Wasserstein distance, to more completely characterize the texture structure. The originality of our contribution is threefold. First, we demonstrate that amplitude information yields statistically significant gains in classification. Second, relying on only eight features, our proposal achieves high accuracy (≈90%) on 28 texture classes, while remaining interpretable via SHapley Additive exPlanations (SHAP) analysis. Third, feature extraction is orders of magnitude faster than the filter bank or deep learning pipelines, making the method particularly attractive in domains where

22 October 2026 09:21:48

computational resources are limited or model transparency is required.

I. INTRODUCTION

Image classification is important across many domains. In particular, texture recognition and discrimination have applications in fields such as remote sensing,[1] geology,[2] geospatial imaging,[3] crystallographic image analysis,[4] and textile defect detection.[5]

The Bandt and Pompe (BP) symbolization technique[6] has been widely used to compute information-theoretic quantifiers for characterizing the dynamics of complex systems. Originally developed for time series, it was later extended to image analysis.[7–9] A central challenge in this context is defining an appropriate strategy for reading an image. Early approaches treated the image as a matrix and analyzed it row by row or through submatrix constructions.[9] Subsequently, alternative traversal strategies have been proposed to map two-dimensional data onto one-dimensional time series. In particular, Chagas *et al.*[3] and Bariviera *et al.*[10] propose the Hilbert curve as a scanning method that eliminates directional bias, preserves spatial locality, and enables the extraction of information-theoretic quantifiers.

More recently, combining Hilbert-based scanning with permutation entropy, permutation statistical complexity, and the Fisher information measure (FIM) has been shown to provide an effective characterization of synthetic images with controlled correlations and real-world textures.[10] Additionally, Hilbert curves have recently been used to uncover spatial dependencies within rectangular data structures.[11] Despite these advances, a key limitation of the original BP framework is that it relies solely on ordinal information, disregarding amplitude variations. To address this limitation, we incorporate two amplitude-sensitive quantifiers: weighted permutation entropy (WPE)[12] and amplitude-aware permutation entropy (AAPE).[13] Both measures explicitly account for amplitude information and have been shown to outperform standard permutation entropy in classification tasks involving signals with non-stationary amplitude dynamics.[14] We also incorporate the Wasserstein (Earth mover's) distance to quantify dissimilarities between probability distributions. Thus, the aim of this paper is to assess whether classification based on two information-theoretic quantifiers (namely, permutation entropy and permutation statistical complexity) achieves satisfactory accuracy or whether additional metrics are required.

In this paper, we propose a two-step information-theoretic framework for texture classification. First, as in Refs. 3 and 15, each image is converted into a one-dimensional time series via a Hilbert curve scan of level 9. Second, eight complementary scalar quantifiers are extracted from the resulting series and used to train a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel, which have proven effective for function approximation and classification in high-dimensional spaces.[16,17]

To ensure the validity of the results, we compute several performance metrics such as accuracy, recall, and the Matthews correlation coefficient.[18] Furthermore, we use the McNemar test[19] to compare feature sets and determine whether observed performance variations are statistically significant or attributable to noise in the training data.[20] Finally, in alignment with the recent literature, we include SHapley Additive exPlanations (SHAP) as a global interpretability method. This allows for the assessment of each quantifier's individual contribution to the final prediction.[21,22]

The remainder of this paper is organized as follows: Sec. II briefly outlines the methodology employed in this study, while Sec. III describes the dataset and statistical metrics used. Section IV presents a detailed analysis of the results. Finally, Sec. V summarizes the primary conclusions and suggests directions for future research.

II. METHODS

A. Hilbert curve scanning

A digital image of size $2^L \times 2^L$ pixels, for some integer L, is converted into a one-dimensional time series by visiting each pixel exactly once along a *planar Hilbert curve* of level L. The Hilbert curve is a space-filling fractal defined in $\mathbb{R}^n$ whose Hausdorff dimension is equal to n.[23] At level L, it partitions the unit square into $2^L \times 2^L$ sub-squares and traces a path that visits every sub-square exactly once. This path is obtained as the limit of a sequence of polygonal curves of increasing length. It is constructed recursively through a fixed set of rotation and reflection rules applied to a base motif. Figure 1 illustrates four levels of the Hilbert curve.

The critical property exploited here is *locality preservation*, i.e., two close points in a one-dimensional arrangement correspond to two close points in the two-dimensional arrangement before unfolding (the converse is not true).[23,24] This property is not shared by row-wise or column-wise scans, which introduce a preferred scanning direction and disrupt spatial correlations between pixels in different rows. The Hilbert curve avoids directional bias because its

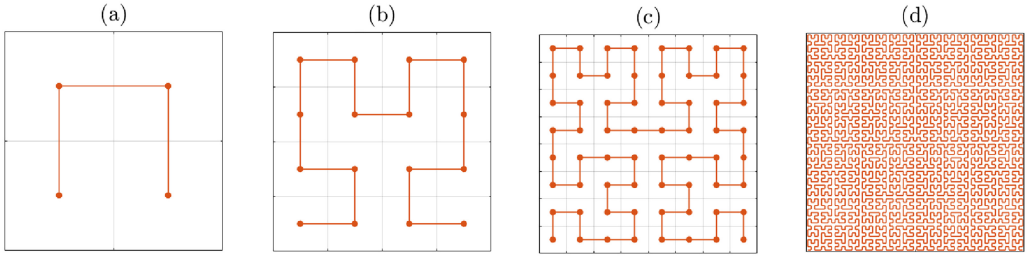


FIG. 1. Levels of the planar Hilbert curve. (a) $L = 1$, (b) $L = 2$, (c) $L = 3$, and (d) $L = 6$.

22 October 2026 09:21:48

© Author(s) 2026

recursive construction does not favor any particular direction. For a comparison of this scanning method with the standard 2×2 pixels symbolization, see Ref. 10.

The pixel intensity values extracted along the curve define the time series $\mathscr{S}(t) = \{x_t\}_{1 \le t \le T}$, with $T = 2^{2L}$, which is analyzed by the information-theoretic quantifiers described below. Although the real-world images are usually rectangles of arbitrary size, in this work, images with $T > 2^L \times 2^L$ are cropped to the largest power-of-two side length prior to scanning, for simplicity. To handle cases of nonsquare images, it can be considered some extensions involving constructions of modified Hilbert curves (see, e.g., Refs. 25 and 11 and the GitHub project by Jakub Červený[26]).

B. Bandt and Pompe symbolization

Let $\mathscr{S}(t) = \{x_t\}_{1 \le t \le T}$ be the time series obtained from the Hilbert scan, let $D > 1$ be the *embedding dimension*, and let $\tau \in \mathbb{N}$ be the *embedding delay*. Following Ref. 6, a sliding window of length D is placed at each position t to form the subsequence,

$$\boldsymbol{w}_t = \left(x_t, x_{t+\tau}, \ldots, x_{t+(D-1)\tau}\right). \tag{1}$$

Each window is mapped to its *ordinal pattern*: the permutation $\pi \in \mathscr{S}_D$ of $\{1, 2, \ldots, D\}$ that sorts the components of $\boldsymbol{w}_t$ in ascending order. There are $D!$ distinct ordinal patterns. The relative frequency with which pattern π_i appears across all $T - (D-1)\tau$ windows defines the *ordinal-pattern probability distribution,*

$$\begin{aligned} P &= \{p(\pi_i), i = 1, \ldots, D!\}, \\ p(\pi_i) &= \frac{\#\{t \le T - (D-1)\tau : \boldsymbol{w}_t \text{has type } \pi_i\}}{T - (D-1)\tau}. \end{aligned} \tag{2}$$

The Bandt–Pompe approach[6] is invariant under monotone nonlinear transformations of the signal, requires no prior knowledge of the underlying dynamics, and is applicable to any time series satisfying that the resulting ordinal-pattern series is stationary (see Refs. 27 and 28 for a discussion).

C. Ordinal-pattern orderings

Several information-theoretic quantifiers, which are defined below, depend on the order in which the $D!$ patterns are indexed. This is because they measure the properties of consecutive differences in the probability distribution.

In this work, two orderings are employed. The first is lexicographic (also known as Lehmer) ordering,[29] where patterns are ranked in ascending lexicographic order. This ordering was adopted in the original formulation of the Fisher information measure for time series[30,31] and is the standard choice in the literature. However, since images can hide more complex structures, we also employ the colexicographic ordering, where patterns are sorted by comparing their elements from the *last* position to the first. In this way, the most recently observed value in the embedding window is the primary sorting key. The colexicographic ordering assigns the most recent observation the greatest weight in the notion of pattern adjacency, providing a complementary perspective to the lexicographic ordering for quantifiers that capture gradient properties of P.

For illustration, consider $D = 3$, and then, we have $D! = 6$ ordinal patterns. We can arbitrarily label them: $\pi_1 = (2, 0, 1)$, $\pi_2 = (0, 2, 1)$, $\pi_3 = (1, 0, 2)$, $\pi_4 = (2, 1, 0)$, $\pi_5 = (0, 1, 2)$, and $\pi_6 = (1, 2, 0)$. The lexicographic (Lehmer) ordering is $\{\pi_5, \pi_2, \pi_3, \pi_6, \pi_1, \pi_4\}$ and the colexicographic ordering is $\{\pi_4, \pi_6, \pi_1, \pi_2, \pi_3, \pi_5\}$.

Whereas global quantifiers (permutation entropy, statistical complexity, weighted permutation entropy, and amplitude-aware permutation entropy) are ordering-invariant, locally sensitive quantifiers such as the Fisher information measure and the Wasserstein distance depend on the ordering.

D. Permutation entropy

Following Ref. 6, the *permutation entropy* of the ordinal pattern distribution P is the normalized Shannon entropy,[32]

$$\mathscr{H}[P] = \frac{S[P]}{S_{\max}} = -\frac{\sum_{i=1}^{D!} p(\pi_i) \ln p(\pi_i)}{\ln(D!)}, \tag{3}$$

where the denominator $\ln(D!)$ is the entropy of the uniform distribution over all $D!$ patterns and ensures $\mathscr{H} \in [0, 1]$. When $\mathscr{H} \approx 1$, the ordinal patterns are nearly equiprobable, indicating a texture close to spatial randomness. When $\mathscr{H} \approx 0$, a small number of patterns dominates, indicating high spatial regularity.

E. Weighted permutation entropy

Standard permutation entropy assigns an equal weight to each embedding window regardless of whether the pixel values in the window span a wide amplitude range or are nearly constant. Fadlallah *et al.*[12] proposed *weighted permutation entropy* (WPE) to address this limitation by weighting each window according to its variance,

$$w(\boldsymbol{w}_t) = \frac{1}{D} \sum_{j=0}^{D-1} \left(x_{t+j\tau} - \bar{x}_t\right)^2, \tag{4}$$

where $\bar{x}_t = D^{-1} \sum_{j=0}^{D-1} x_{t+j\tau}$ is the mean of the window. The weighted probability of pattern π_i is then

$$p_w(\pi_i) = \frac{\sum_{t:\boldsymbol{w}_t \sim \pi_i} w(\boldsymbol{w}_t)}{\sum_t w(\boldsymbol{w}_t)}, \tag{5}$$

and WPE is defined as the normalized Shannon entropy of the weighted distribution,

$$\mathscr{H}_w[P_w] = -\frac{\sum_{i=1}^{D!} p_w(\pi_i) \ln p_w(\pi_i)}{\ln(D!)}. \tag{6}$$

Texture regions with large-amplitude fluctuations contribute more to the distribution than flat, low-variance windows, making WPE sensitive to both the ordinal structure and the amplitude dynamics of the signal. For textures with similar values of $\mathcal{H}$, WPE may differ substantially if one has large-amplitude fluctuations and the other does not. This makes WPE a complementary descriptor to $\mathcal{H}$. This is especially relevant for natural textures, where high-variance regions (e.g., edges, coarse grain boundaries) carry qualitatively different information from smooth homogeneous regions.

22 October 2026 09:21:48

© Author(s) 2026

F. Amplitude-aware permutation entropy

A limitation of WPE is that the variance weight $w(\boldsymbol{w}_t)$ vanishes for embedding windows with constant amplitude, even when a well-defined ordinal pattern exists. As a result, ordinal patterns arising from flat regions do not contribute to the distribution, regardless of their absolute intensity level. Azami and Escudero[13] proposed *amplitude-aware permutation entropy* (AAPE) to overcome this by using a two-component weight that combines the mean absolute amplitude of the window and its mean absolute consecutive difference,

$$A(\boldsymbol{w}_t) = \frac{\alpha}{D}\sum_{j=0}^{D-1}|x_{t+j\tau}| + \frac{\beta}{D-1}\sum_{j=0}^{D-2}|x_{t+(j+1)\tau} - x_{t+j\tau}|, \tag{7}$$

where $\alpha,\ \beta \geq 0$ with $\alpha+\beta=1$ control the relative contribution of each term. The first term captures the average intensity level of the window, while the second term captures its local rate of change, quantifying local signal fluctuations through a discrete first-order finite difference. Following Azami and Escudero,[13] we set $\alpha=\beta=0.5$ throughout this work. The amplitude-aware weighted probability and the AAPE are then defined analogously to Eqs. (5) and (6),

$$\begin{aligned} p_A(\pi_i) &= \frac{\sum_{t:\boldsymbol{w}_t\sim\pi_i} A(\boldsymbol{w}_t)}{\sum_t A(\boldsymbol{w}_t)}, \\ \mathscr{H}_A[P_A] &= -\frac{\sum_{i=1}^{D!} p_A(\pi_i)\ln p_A(\pi_i)}{\ln(D!)}. \end{aligned} \tag{8}$$

Unlike WPE, AAPE assigns a non-zero weight to constant-amplitude windows whenever their mean absolute amplitude is non-zero. Consequently, it captures information from flat regions that would otherwise be discarded. AAPE, therefore, provides complementary information to WPE, particularly in textures where the mean gray level varies spatially.

G. Statistical complexity

Relying solely on Shannon entropy for time-series analysis often fails to capture the underlying structure of complex systems. As Feldman and Crutchfield[33] noted, entropy alone cannot account for their intrinsic organizational structure. Statistical complexity measures are then employed, in some cases, to provide a more comprehensive perspective.

Under this framework, time series featuring a sophisticated interplay of patterns receive higher complexity scores, reflecting a high degree of internal structure. Conversely, systems at the extremes (those that are perfectly ordered or entirely disordered) are assigned lower complexity values. This capacity to distinguish systems based on their level of organizational depth highlights the critical utility of statistical complexity in modern analysis.

To address this, López-Ruiz *et al.*[34] proposed a complexity measure based on the concept of disequilibrium. Originally, disequilibrium was defined using the squared Euclidean distance.[34] Subsequently, in a series of works Lamberti *et al.*[35] and Martín *et al.*[36] redefined this concept, experimenting with different distance metrics, and finding that the Jensen–Shannon divergence is the best suited disequilibrium metric for discriminating stochastic and chaotic processes. Thus, the *permutation statistical complexity* is defined as

$$\mathscr{C}_{JS}[P] = \mathscr{Q}_J[P,P_e]\cdot\mathscr{H}[P], \tag{9}$$

where P_e denotes the uniform distribution ($p_j = 1/D!$ for all j) and $\mathscr{Q}_J$ is the Jensen–Shannon disequilibrium,

$$\mathscr{Q}_J[P,P_e] = \mathscr{Q}_0\cdot\mathscr{J}[P,P_e], \tag{10}$$

with the Jensen–Shannon divergence,[37]

$$\mathscr{J}[P,P_e] = S\left[\frac{P+P_e}{2}\right] - \frac{S[P]}{2} - \frac{S[P_e]}{2}, \tag{11}$$

and the normalization constant,

$$\mathscr{Q}_0 = -2\left[\frac{N+1}{N}\ln(N+1) - 2\ln(2N) + \ln N\right]^{-1}, \tag{12}$$

where $N=D!$. The product structure of Eq. (9) ensures that $\mathscr{C}_{JS}=0$ for both completely random processes ($\mathscr{H}=1$, $\mathscr{Q}_J=0$) and perfectly regular ones ($\mathscr{H}=0$, $\mathscr{Q}_J\to 0$), while reaching its maximum for processes with intermediate structure. The pair $(\mathscr{H},\mathscr{C}_{JS})$ is represented in the *Complexity-entropy Causality Plane* (CECP), first introduced by Rosso *et al.*[38] as a tool for discriminating stochastic from chaotic dynamics. The CECP has been widely used in dynamical systems,[39,40] financial time series,[41,42] and image textures analysis.[8,9]

H. Fisher information measure

Unlike Shannon entropy, which is a global measure of disorder, the *Fisher information measure* (FIM) is a local quantity that captures the gradient content of a probability distribution, i.e., how rapidly adjacent probability varies across the ordered support.[43,44] For a discrete probability distribution $P=\{p_j, j=1,\ldots,N\}$ ordered according to a given permutation σ, the normalized discrete FIM reads

$$\mathscr{F}^{\sigma}[P] = \mathscr{F}_0\sum_{j=1}^{N-1}\left[\sqrt{p_{j+1}} - \sqrt{p_j}\right]^2, \tag{13}$$

where the normalization constant is

$$\mathscr{F}_0 = \begin{cases} 1, & \text{if } p_{j^*}=1 \text{for } j^*=1, \\ & \text{or } j^*=N, \text{ and } p_j=0\ \forall j\neq j^*, \\ 1/2, & \text{otherwise.} \end{cases} \tag{14}$$

This discretization is well-behaved[45] and ensures that $\mathscr{F}\in[0,1]$. A highly ordered system, where the distribution is concentrated at a single pattern, has low entropy ($\mathscr{H}\approx 0$) and high FIM ($\mathscr{F}\approx 1$). In contrast, a uniform distribution yields high entropy ($\mathscr{H}\approx 1$) and low FIM ($\mathscr{F}\approx 0$). Thus, FIM behaves oppositely to Shannon entropy, except for periodic motions.[31] The pair $(\mathscr{H},\mathscr{F})$ constitutes the Fisher-entropy Causality Plane (FECP), introduced by Vignat and Bercher.[46]

Since Eq. (13) depends on the ordering of P, two values are computed per image: $\mathscr{F}^{\text{lex}}$ using lexicographic ordering and $\mathscr{F}^{\text{colex}}$ using colexicographic ordering (Sec. II C). These two values encode different notions of pattern adjacency, capturing complementary aspects of the local correlation structure of the texture.

22 October 2026 09:21:48

© Author(s) 2026

I. Wasserstein distance

The *1-Wasserstein distance* (also known as the Earth mover's distance or Kantorovich–Rubinstein metric) is a measure of dissimilarity between two probability densities over a metric space Ω. In this work, we compute the dissimilarity between the image ordinal pattern distribution P and the uniform distribution P_e, interpreting it as the minimum amount of probability mass that must be transported to transform P into P_e, under a ground metric induced by the ordering σ. For a discrete ordered support of size N with unit distance between consecutive support points, the 1D Wasserstein distance reduces to the ℓ^1 norm of the difference between cumulative distribution functions,[47]

$$\mathscr{W}_1^{\sigma}(P, P_e) = \sum_{k=1}^{N} \left| \mathrm{CDF}_P^{\sigma}(k) - \frac{k}{N} \right|, \tag{15}$$

where $\mathrm{CDF}_P^{\sigma}(k) = \sum_{j=1}^{k} p_{\sigma(j)}$ is the cumulative distribution function of P under ordering σ and k/N is the CDF of the uniform distribution (which is independent of ordering). $\mathscr{W}_1^{\sigma}$ measures how far the ordinal-pattern distribution deviates from equiprobability: values close to zero indicate a nearly uniform distribution (spatially random texture), whereas larger values indicate the concentration of probability mass (structured texture). Since P_e is uniform, its CDF does not depend on σ, but the CDF of P changes with the ordering, making $\mathscr{W}_1^{\sigma}$ an ordering-sensitive quantifier, complementary to $\mathscr{F}^{\sigma}$. As for the FIM, two values are computed per image: $\mathscr{W}_1^{\mathrm{lex}}$ and $\mathscr{W}_1^{\mathrm{colex}}$. The use of Wasserstein distance in image analysis can be traced back to Peleg *et al.*,[48] who introduced a "match-distance" framework to quantify the minimal transport cost between gray-level distributions of images, which is closely related to the Wasserstein formulation.

J. Classification model

Texture classification was performed using a support vector machine (SVM) with a radial basis function (RBF) kernel, defined as

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right). \tag{16}$$

A Support Vector Machine (SVM) is a supervised machine learning algorithm designed for classification tasks. It finds an optimal hyperplane that maximizes the distance between different classes within an N-dimensional space. Originally introduced by Vapnik *et al.*[17], SVMs and their variants have become foundational tools for addressing complex classification problems.[1,49]

Feature standardization (zero mean and unit variance) was applied within each training fold. Multi-class classification was handled using a one-vs-one strategy, yielding $K(K-1)/2$ binary classifiers for K classes. Posterior probabilities were estimated using Platt scaling to enable probabilistic metrics.

Only the RBF-kernel was considered. Linear SVMs were excluded after preliminary experiments (not reported in this paper) consistently showed inferior performance, which is expected given the nonlinear structure of entropy-complexity plane and transport-based descriptors used in this study.

Various classification algorithms (e.g., random forests or XGBoost) could be employed; however, the optimal choice often depends on specific dataset characteristics and feature dimensionality.[1] The primary objective of this study, however, is to evaluate the impact of introducing additional metrics into the classification process, rather than to perform a comparative benchmark of different algorithms. Consequently, we have utilized SVM as a consistent baseline. This approach ensures that changes in performance can be attributed to the proposed metrics, providing readers with clearer insights and more focused conclusions.

III. DATA AND STATISTICAL PERFORMANCE METRICS

Experiments were conducted on the Kylberg Texture Dataset (v. 1.0).[50] The dataset comprises 28 distinct texture classes representing a broad range of natural and manufactured surface types, including fabrics (blanket, scarf, rug), food items (rice, lentils, oatmeal, linseeds, sesame seeds, pearl sugar), floor and ceiling materials (canvas, stone, stone slab, wall), and organic materials (grass, sand). Each class contains ten grayscale images of size 512×512 pixels, yielding a total of $M = 280$ images. A sample of each class is displayed in Fig. 2.

To evaluate the performance of the classification model, several statistical metrics are employed to provide a comprehensive assessment of its predictive power. Although these metrics are well established within the machine learning community, formal definitions are provided below to ensure the article remains self-contained.

Let K denote the number of classes and n_k be the number of images belonging to class k, with $M = \sum_{k=1}^{K} n_k$. For each class k, we define the following quantities from the confusion matrix: TP_k (true positives), FP_k (false positives), FN_k (false negatives), and TN_k (true negatives), where the subscript refers to treating class k as the positive class in a one-vs-rest decomposition:

$$\mathrm{TP}_k = C_{kk},$$

$$\mathrm{FP}_k = \sum_{l \neq k} C_{lk} \quad \text{(other classes predicted as } k\text{)},$$

$$\mathrm{FN}_k = \sum_{l \neq k} C_{kl} \quad \text{(class } k \text{predicted as other classes)},$$

$$\mathrm{TN}_k = N - \mathrm{TP}_k - \mathrm{FP}_k - \mathrm{FN}_k.$$

A. Accuracy

Overall accuracy is the fraction of images assigned to their correct class,

$$\mathrm{ACC} = \frac{\sum_{k=1}^{K} \mathrm{TP}_k}{M}. \tag{17}$$

B. Per-class precision

Precision for class k measures the fraction of images *predicted as* class k that actually belong to class k,

$$\mathrm{P}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}. \tag{18}$$

A low value of P_k indicates that the classifier frequently assigns images from other classes to class k (many false positives).

22 October 2026 09:21:48

© Author(s) 2026

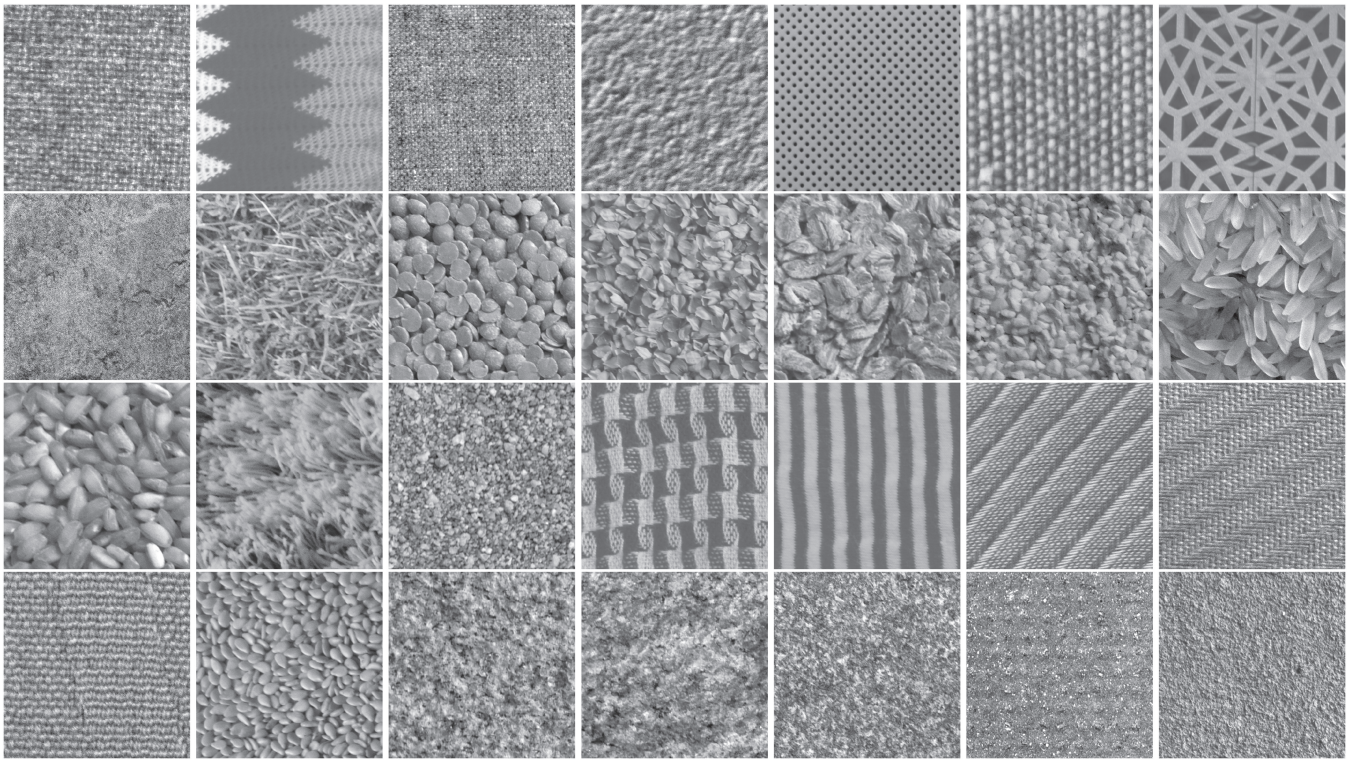


FIG. 2. Sample of each class of the Kylberg Dataset v. 1.0.

C. Per-class recall

Recall (also called sensitivity or true positive rate) for class k measures the fraction of images that belong to class k that are correctly identified,

$$\mathrm{R}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}. \tag{19}$$

A low value of R_k indicates that the classifier misses many images from class k (many false negatives).

D. Per-class F1 score

The F1 score for class k is the harmonic mean of precision and recall, combining both error types into a single class-level summary,

$$\mathrm{F1}_k = \frac{2\mathrm{P}_k\mathrm{R}_k}{\mathrm{P}_k + \mathrm{R}_k} = \frac{2\mathrm{TP}_k}{2\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k}. \tag{20}$$

$\mathrm{F1}_k = 0$ when either precision or recall is zero, and $\mathrm{F1}_k = 1$ when both are perfect. Per-class F1 scores are reported for the in-sample evaluation to identify which specific texture categories are most difficult to discriminate.

E. Macro-averaged F1 score

The macro-averaged F1 score is the unweighted arithmetic mean of per-class F1 scores across all K classes,

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{F1}_k, \tag{21}$$

giving equal weight to every class regardless of its size. This is the primary F1 metric reported in all evaluation tables.

F. Matthews correlation coefficient

The Matthews correlation coefficient (MCC)[18] is a single scalar summary of the full $K \times K$ confusion matrix. For multi-class problems, it is defined as[51]

$$\mathrm{MCC} = \frac{\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{m=1}^{K} C_{kk}C_{lm} - C_{kl}C_{mk}}{\sqrt{\left(\sum_{k=1}^{K}\sum_{l=1}^{K} C_{kl}^2\right) - \sum_{k=1}^{K}\left(\sum_{l=1}^{K} C_{kl}\right)^2} \cdot \sqrt{\left(\sum_{k=1}^{K}\sum_{l=1}^{K} C_{kl}^2\right) - \sum_{k=1}^{K}\left(\sum_{l=1}^{K} C_{lk}\right)^2}}, \tag{22}$$

22 October 2026 09:21:48

© Author(s) 2026

where C_{kl} is the (k, l) entry of the confusion matrix (the number of images from class k predicted as class l). The MCC takes values in $[-1, 1]$: a value of $+1$ indicates perfect prediction, 0 indicates prediction no better than random, and -1 indicates total disagreement between predictions and true labels. The MCC is considered one of the most informative single-number summaries for multi-class problems because it accounts for all cells of the confusion matrix simultaneously and is not inflated by class imbalance.[52]

G. Macro-averaged ROC–AUC

The area under the receiver operating characteristic curve (ROC–AUC) measures a classifier's ability to rank images of the positive class above images of the negative class without committing to a fixed decision threshold. For K classes, the macro-averaged one-vs-rest ROC–AUC is

$$\mathrm{AUC}_{\mathrm{macro}} = \frac{1}{K}\sum_{k=1}^{K}\int_0^1 \mathrm{TPR}_k\left(\mathrm{FPR}_k^{-1}(t)\right)\mathrm{d}t, \tag{23}$$

where $\mathrm{TPR}_k = \mathrm{TP}_k/(\mathrm{TP}_k + \mathrm{FN}_k)$ and $\mathrm{FPR}_k = \mathrm{FP}_k/(\mathrm{FP}_k + \mathrm{TN}_k)$ are the true and false positive rates for class k at a given decision threshold. The AUC ranges from 0 to 1; a value of 1 denotes perfect discrimination and 0.5 corresponds to random ranking. Unlike the metrics above, the ROC–AUC evaluates the quality of the predicted probability scores rather than hard class assignments, making it a useful complement when the decision threshold is not fixed in advance. Probability estimates were obtained via Platt scaling,[16] which calibrates the SVM decision function values into proper class probabilities.

These metrics form a hierarchy of informational depth. While accuracy reflects overall correctness, it neglects class distribution. For a more robust assessment, the F1 score and MCC are preferred, as they jointly penalize false positives and false negatives.[52] Conversely, ROC–AUC measures ranking ability independent of specific thresholds, capturing performance nuances that accuracy might miss. Together, these metrics provide a comprehensive evaluation that no single value can offer.

IV. RESULTS

All images were scanned using a Hilbert curve of level 9, generating a one-dimensional time series of $2^{9\times 2} = 262.144$ pixel values per image. Information-theoretic and transport-based quantifiers were then computed with embedding dimension $D = 6$ and delay $\tau = 1$, following Ref. 10.

Figure 3 displays the complexity-entropy causality plane, the Fisher-entropy plane, the complexity-Wasserstein plane, and the complexity-weighted permutation entropy plane. Although the CECP partially discriminates certain classes, it exhibits significant overlap among several others, preventing accurate classification. Additionally, other information-theoretic and transport-based quantifiers capture also non-trivial characteristics that could eventually help with the classification task since the points on the plane are more clearly separated.

Therefore, and in agreement with the original aim of this paper, feature sets were compared:

TABLE I. In-sample classification metrics for the Kylberg dataset. Both feature sets used $C = 100$, $\gamma = \texttt{scale}$.

Feature set	Accuracy	F1 (macro)	MCC	ROC–AUC
Set-1 $(\mathcal{H}, C)$	0.618	0.586	0.607	0.978
Set-2 (all eight)	0.957	0.959	0.956	0.998
Δ (set-2–set-1)	+0.339	+0.373	+0.349	+0.020
Δ%	+54.9%	+63.6%	+57.6%	+2.1%

- **Set-1** $[\mathcal{H}, \mathscr{C}_{JS}]$: permutation entropy and statistical complexity, the classical complexity-entropy causality plane baseline.[9,38]
- **Set-2** [all eight features]: $\mathcal{H}$, $\mathscr{C}_{JS}$, $\mathscr{F}^{\mathrm{lex}}$, $\mathscr{F}^{\mathrm{colex}}$, $\mathscr{W}_1^{\mathrm{lex}}$, $\mathscr{W}_1^{\mathrm{colex}}$, $\mathscr{H}_w$ (WPE) and $\mathscr{H}_A$ (AAPE).

Classification was performed using an RBF-kernel SVM with hyperparameters tuned by exhaustive grid search ($C \in \{0.1, 1, 10, 100\}$, $\gamma \in \{\texttt{scale}, \texttt{auto}, 0.01, 0.001\}$) via stratified fivefold inner cross-validation scored by balanced accuracy.

A. In-sample evaluation

Table I reports the in-sample classification metrics for both feature sets. These results represent optimistic upper bounds obtained by fitting and scoring the classifier on the complete 280-image dataset; they are included as a measure of the features discriminative capacity rather than as estimates of generalization performance.

The improvement from set-1 to set-2 is substantial across all metrics. Accuracy rises from 0.618 to 0.957, a gain of 54.9%. The MCC, the most reliable single metric for a multi-class problem with imbalanced class sizes, improves by 57.6%. The ROC–AUC improvement is modest (+2.1%, from 0.978 to 0.998) because the class-ranking ability of the SVM is already high under set-1, meaning that the two classical features capture enough (in-sample) inter-class separation to rank classes correctly, but are insufficient for reliable hard assignment. ROC–AUC measures whether the correct class receives a higher predicted probability than competing classes. This is a task that $\mathscr{H}$ and $\mathscr{C}_{JS}$ accomplish reasonably well since different texture types occupy distinguishable regions of the CECP. What $\mathscr{H}$ and $\mathscr{C}_{JS}$ cannot provide is sufficient probability margin to make reliable hard assignments: the predicted probabilities of the correct and nearest-competing class are often separated by only a few percentage points (due to the overlapping classes in the CECP), leading to frequent misclassification while still contributing positively to the rank-based AUC calculation.

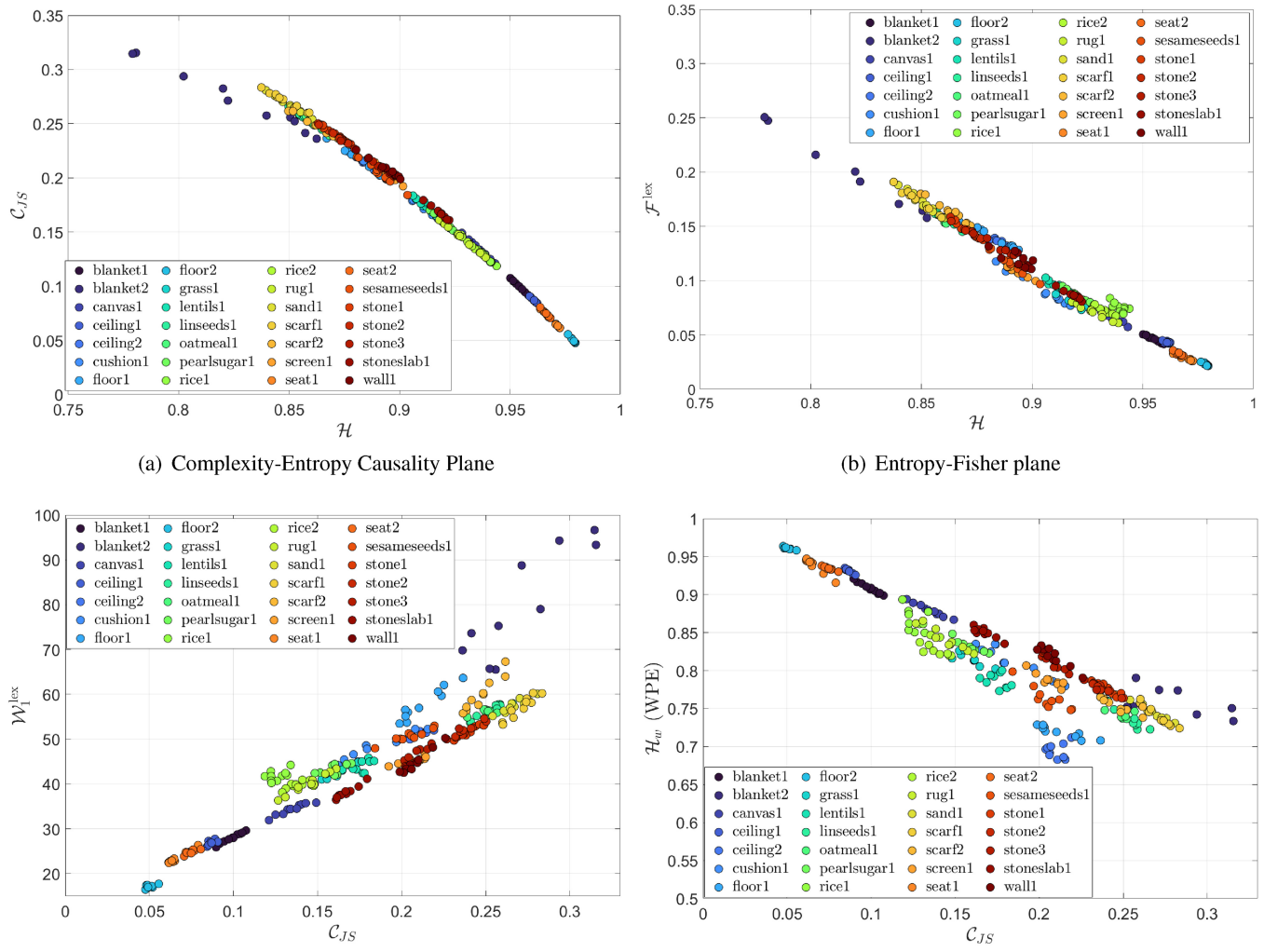
1. Per-class analysis

Table II presents the per-class precision, recall, and F1 score for both feature sets under the in-sample fit. Several patterns are noteworthy.

Under set-1, classes 6 (cushion1) and 16 (rug1) achieve zero precision and recall, the classifier never assigns any image to these classes because the $(\mathcal{H}, \mathscr{C}_{JS})$ values of cushion and rug textures overlap entirely with neighboring classes. Class 19 (scarf2) is almost indistinguishable from other scarf-type textures under set-1, achieving F1= 0.133. The most difficult classes under set-2 are 24 (stone1)

22 October 2026 09:21:48

© Author(s) 2026



(a) Complexity-Entropy Causality Plane

(b) Entropy-Fisher plane

(c) Complexity-Wasserstein Plane

(d) Complexity-Weighted Permutation Entropy Plane

FIG. 3. Bivariate plots of selected information theory and transport-based quantifiers, for $D = 6$ and $\tau = 1$.

and 25 (stone2), which are two varieties of the same material; these achieve an F1 score of 0.737 and 0.640, respectively. The confusion between stone1 and stone2 is reflected in the confusion matrix discussed below.

B. Out-of-sample evaluation

1. Stratified k-fold cross-validation

To ensure a robust evaluation, we employed stratified k-fold cross-validation, which partitions the dataset into k disjoint subsets of approximately equal size. Unlike standard splits, stratification ensures that each fold preserves the original class proportions. The model is then trained and evaluated k times: in each round, one fold serves as the test set and the remaining $k-1$ folds form the training set. Every image appears in the test set exactly once across the k rounds, so the procedure produces k independent performance estimates, whose mean and standard deviation are reported. In this paper, we use $k = 5$ and $k = 7$. For the Kylberg dataset ($N = 280$, $K = 28$), a fivefold cross-validation assigns 56 images (two per class) to each test phase.

Table III presents the results of the stratified fivefold and sevenfold cross-validation. In both cases, every image is assigned to the test set exactly once, ensuring an unbiased evaluation of the entire dataset.

Under fivefold cross-validation, set-2 achieves an accuracy of 0.904 ± 0.024, a strong result for a 28-class problem with only ten training examples per class per fold (i.e., eight training images per class in the 80% training partition). The standard deviations are small (≤ 0.031 for all set-2 metrics), indicating stable performance across different fold assignments. The sevenfold cross-validation

22 October 2026 09:21:48

© Author(s) 2026

TABLE II. Per-class precision, recall, and F1 score for set-1 and set-2 (in-sample, Kylberg dataset, 28 classes, 10 images per class). Classes where set-2 achieves perfect F1 = 1.0 are omitted for brevity; these are classes 1–10, 13–16, 18–20, 23, and 27.

	Set-1			Set-2		
Class	Precision	Recall	F1 score	Precision	Recall	F1 score
6 (cushion1)	0.000	0.000	0.000	1.000	1.000	1.000
7 (floor1)	1.000	0.300	0.462	1.000	1.000	1.000
9 (grass1)	0.500	0.300	0.375	1.000	1.000	1.000
11 (linseeds1)	0.625	0.500	0.556	1.000	0.800	0.889
12 (oatmeal1)	0.500	0.800	0.615	0.833	1.000	0.909
16 (rug1)	0.000	0.000	0.000	1.000	1.000	1.000
17 (sand1)	0.583	0.700	0.636	1.000	0.900	0.947
19 (scarf2)	0.200	0.100	0.133	1.000	1.000	1.000
21 (seat1)	1.000	0.800	0.889	1.000	0.900	0.947
22 (seat2)	0.833	1.000	0.909	0.909	1.000	0.952
24 (stone1)	0.625	0.500	0.556	0.778	0.700	0.737
25 (stone2)	0.375	0.600	0.462	0.533	0.800	0.640
26 (stone3)	0.333	0.100	0.154	1.000	0.800	0.889
28 (wall1)	0.800	0.800	0.800	1.000	0.900	0.947
Macroaverage	0.609	0.618	0.586	0.966	0.957	0.959

yields slightly lower accuracy (0.893 ± 0.022) for set-2, reflecting the reduced training set size (approximately seven images per class in each fold's training partition).

Set-1 achieves an accuracy of approximately 0.51–0.54 under both fold configurations, which is substantially better than random chance ($1/28 \approx 0.036$) but far inferior to set-2. The high ROC–AUC values for set-1 ($\approx$0.968) confirm that the $(\mathcal{H}, \mathscr{C}_{JS})$ features provide good class-ranking ability, but the hard assignment boundary learned from only two features is insufficient for reliable 28-class discrimination.

2. Repeated random splits

While k-fold cross-validation is more data-efficient than a single split, ensuring that no data are excluded from evaluation, repeated random splits (e.g., 80/20) provide a more granular view of performance. By generating a larger pool of independent test scores, repeated splits enable rigorous statistical validation using the Wilcoxon signed-rank test. In this paper, we performed 20 independent stratified random splits using 70/30 and 80/20 proportions between training and test sets. The results are reported in Table IV. All eight Wilcoxon signed-rank tests comparing set-2 against set-1 yield p-value <0.001 across all 20 runs, confirming that the performance gains are statistically significant and consistent.

The 80/20 split yields slightly higher accuracy than the 70/30 split for set-2 (0.907 vs 0.896), which is consistent with the expectation that more training data improve classification in this small-sample regime. The standard deviations for set-2 across the 20 runs are very small ($\leq$0.032), indicating robust and consistent performance. For set-1, the standard deviations are larger, reflecting the greater sensitivity of a two-feature classifier to the specific composition of the training partition.

TABLE III. Stratified k-fold cross-validation results for the Kylberg dataset (mean $\pm$ std over k folds).

k	Set	Accuracy	F1 (macro)	MCC	ROC–AUC
5	Set-1	0.511 ± 0.047	0.460 ± 0.055	0.497 ± 0.049	0.969 ± 0.007
	Set-2	0.904 ± 0.024	0.901 ± 0.026	0.901 ± 0.025	0.996 ± 0.001
	Δ	+0.393	+0.441	+0.404	+0.028
7	Set-1	0.532 ± 0.083	0.472 ± 0.082	0.521 ± 0.086	0.968 ± 0.007
	Set-2	0.893 ± 0.022	0.871 ± 0.031	0.890 ± 0.023	0.996 ± 0.001
	Δ	+0.361	+0.399	+0.369	+0.028

TABLE IV. Out-of-sample (OOS) performance on the Kylberg dataset (mean $\pm$ std. dev. over 20 independent stratified splits; Wilcoxon signed-rank test p-values for set-2 vs set-1).

Split	Metric	Set-1	Set-2	Δ	p-value
70/30	Accuracy	0.520 ± 0.041	0.896 ± 0.021	+0.377	<0.001
	F1 (macro)	0.476 ± 0.042	0.894 ± 0.022	+0.418	<0.001
	MCC	0.506 ± 0.042	0.894 ± 0.022	+0.388	<0.001
	ROC–AUC	0.967 ± 0.006	0.995 ± 0.002	+0.028	<0.001
80/20	Accuracy	0.514 ± 0.060	0.907 ± 0.031	+0.393	<0.001
	F1 (macro)	0.461 ± 0.058	0.899 ± 0.032	+0.438	<0.001
	MCC	0.501 ± 0.063	0.905 ± 0.032	+0.404	<0.001
	ROC–AUC	0.967 ± 0.004	0.996 ± 0.001	+0.028	<0.001

C. McNemar's test

McNemar's test is a non-parametric statistical hypothesis test used to compare the predictive performance of two classifiers evaluated on the same dataset. The null hypothesis states that both classifiers have the same probability of making a correct prediction.[19,20] Table V reports McNemar's exact test results comparing set-1 and set-2 predictions on the same images across all evaluation contexts. The test focuses on the discordant pairs, i.e., images that one feature set classifies correctly, but the other does not.

The results are unambiguous. Across all evaluation contexts, the number of images corrected by set-2 relative to set-1 (column c) vastly exceeds the number lost (column b). In the pooled 70/30 OOS data, set-2 corrects 38.3% of the predictions that set-1 gets wrong, while losing only 0.7%, a ratio of approximately 59:1. The

TABLE V. McNemar's exact test: set-1 vs set-2. Columns a–d correspond to (a) both correct, (b) set-1 correct/set-2 wrong, (c) set-1 wrong/set-2 correct, and (d) both wrong.

Context	n	a (%)	b (%)	c (%)	d (%)	p-value
In-sample	280	60.4	1.4	35.4	2.9	<0.001
Fivefold cross-validation	280	50.4	0.7	40.0	8.9	<0.001
Sevenfold cross-validation	280	52.5	0.7	36.8	10.0	<0.001
Pooled OOS 70/30	1680	51.3	0.7	38.3	9.7	<0.001
Pooled OOS 80/20	1120	50.7	0.7	40.0	8.6	<0.001

22 October 2026 09:21:48

© Author(s) 2026

TABLE VI. Noise robustness: mean accuracy ± std (30 realizations) for the Kylberg dataset. Retention is the percentage relative value to the accuracy value without noise.

σ	Set-1 ($\mathscr{H}, \mathscr{C}_{JS}$)	Retention (%)	Set-2 (all 8)	Retention (%)
0.000	0.618 ± 0.000	…	0.957 ± 0.000	…
0.010	0.613 ± 0.009	99.2	0.958 ± 0.005	100.1
0.025	0.586 ± 0.015	94.9	0.955 ± 0.008	99.8
0.050	0.509 ± 0.022	82.4	0.933 ± 0.010	97.5
0.100	0.401 ± 0.026	65.0	0.836 ± 0.019	87.3
0.200	0.286 ± 0.022	46.3	0.642 ± 0.032	67.1
0.400	0.188 ± 0.022	30.4	0.423 ± 0.027	44.2

corresponding p-values lead us to reject the null hypothesis at the 1% significance level. This confirms that the two feature sets make fundamentally different errors and that the extended set is systematically superior. This is not a marginal improvement driven by one or two lucky test partitions; it holds across all 20 independent random splits, which amounts to strong evidence of genuine generalization.

D. Noise robustness

To evaluate sensitivity to feature perturbations, we added independent Gaussian noise to each feature with a standard deviation σ proportional to its empirical feature variance. Models trained on clean data were evaluated using noisy feature realizations in 30 repetitions at each noise level. Accuracy was reported as a function of σ. This experiment evaluates decision boundary stability under feature perturbations, rather than robustness to real-world acquisition noise.

Table VI displays the noise robustness of both feature sets on the Kylberg dataset. For each noise level σ (expressed as a fraction of each feature's standard deviation), Gaussian noise $\mathcal{N}(0, \sigma^2 \hat{\sigma}_j^2)$ was added to the test features while the model was trained on clean data. Accuracy was averaged over 30 independent noise realizations. Table VI shows that, at the reference noise level of $\sigma = 0.05$ (5% of each feature's natural variability), set-2 retains 97.5% of its clean-data performance (0.933 vs 0.957), while set-1 retains only 82.4% (0.509 vs 0.618). This difference reflects the fact that set-2's eight-dimensional feature vector is inherently more redundant; perturbation of any single feature has a smaller relative impact on the classification boundary. At $\sigma = 0.10$, set-2 degrades to 0.836, which is still substantially above set-1's clean-data baseline of 0.618, while set-1 falls to 0.401, approaching the level of chance. Therefore, the classification scheme performs is robust even under small noise perturbations.

E. SHAP feature importance

In this section, we scrutinize which features contribute the most to the classification task. For this purpose, we apply SHapley Additive exPlanations (SHAP),[21,22] a model-agnostic framework rooted in cooperative game theory that quantifies the contribution of each feature to individual predictions by computing its average marginal contribution across all possible feature subsets. Unlike global importance measures such as permutation importance or mean decrease impurity, SHAP provides locally faithful attributions

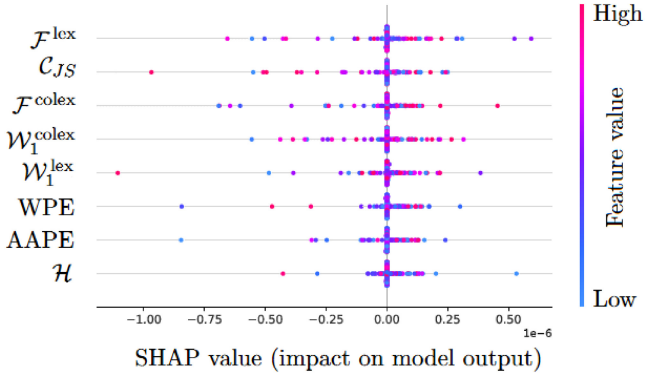


FIG. 4. SHAP summary plot (RBF-kernel SVM, class-averaged).

that satisfy the Shapley axioms of efficiency, symmetry, and additivity, offering stronger theoretical guarantees than gradient-based or perturbation-based alternatives.

Figure 4 shows the SHAP beeswarm plot for the eight-feature RBF-kernel SVM trained on the Kylberg dataset. Each dot represents one image; its horizontal position indicates the SHAP value (the contribution of that feature to the model output for that image), and its color encodes the feature value, ranging from blue (low) to red (high). Features are arranged vertically from top to bottom in decreasing order by the mean absolute value of the class-averaged SHAP contributions. It is important to note that this plot averages the signed SHAP values across all 28 classes before plotting: positive contributions for one class can partially cancel negative contributions for another, which is why most dots cluster near zero and the scale of the x axis is on the order of 10^{-6}. To avoid this cancellation effect, the global importance reported in Table VII is computed as the average across classes of the mean absolute SHAP values, which preserves feature influence regardless of sign.

Three patterns emerge from joint reading of Fig. 4 and Table VII. First, $\mathscr{F}^{\text{lex}}$ occupies the top position in the beeswarm because it retains a consistent sign even after cross-class averaging: high values of $\mathscr{F}^{\text{lex}}$ (red dots) tend to appear on the right side of the plot, indicating a positive contribution to the class-averaged model output, while low values (blue dots) appear on the left. This feature is most critical for class 6 (cushion1, mean $|\text{SHAP}| = 0.0273$), which exhibits a distinctive local gradient structure in its ordinal pattern distribution that the Fisher measure captures and entropy-based features miss entirely. Second, WPE ranks first in the global importance table despite appearing lower in the beeswarm, precisely because its contributions are large in magnitude but of mixed sign across classes: it strongly helps discriminate grain-like textures such as lentils (class 10, 0.0380) but contributes in the opposite direction for other textures, so the averaging largely cancels these effects in the beeswarm. This confirms that WPE is the most discriminative single feature globally, but its influence is highly class-specific rather than uniformly directional. Third, permutation entropy $\mathscr{H}$ and statistical complexity $\mathscr{C}_{JS}$ consistently rank last in the global table (ranks 8 and 6, respectively) and show the narrowest spread

22 October 2026 09:21:48

© Author(s) 2026

TABLE VII. Global and selected per-class SHAP importance (mean |SHAP| value) for the Kylberg dataset. Features are ranked by global importance (descending). Per-class columns show the five most informative classes for each feature.

Feature	Global rank	Global mean |SHAP|	Most informative class	Class mean |SHAP|
WPE	1	0.0130	Class 10 (lentils1)	0.0380
AAPE	2	0.0096	Class 20 (screen1)	0.0245
$\mathscr{F}^{\text{colex}}$	3	0.0087	Class 6 (cushion1)	0.0254
$\mathscr{F}^{\text{lex}}$	4	0.0082	Class 6 (cushion1)	0.0273
$\mathscr{W}_1^{\text{lex}}$	5	0.0081	Class 6 (cushion1)	0.0250
$\mathscr{C}_{JS}$	6	0.0076	Class 14 (rice1)	0.0158
$\mathscr{W}_1^{\text{colex}}$	7	0.0076	Class 2 (blanket2)	0.0096
$\mathcal{H}$	8	0.0067	Class 14 (rice1)	0.0137

in the beeswarm. Both retain modest class-specific importance. For example, $\mathscr{C}_{JS}$ and $\mathscr{H}$ are most informative for class 14 (rice1), suggesting that the global disorder of the rice texture is its primary distinguishing characteristic. However, neither feature provides sufficient discrimination on its own, confirming that the CECP, while informative, must be extended with local and amplitude-sensitive quantifiers to achieve reliable 28-class texture classification.

V. CONCLUSIONS

Three primary conclusions emerge from this study regarding the efficacy of Hilbert curve-based information quantifiers. First, the eight-feature set achieves substantially superior performance across all evaluation schemes, with accuracy improvements of 33–39 percentage points. This improvement is statistically robust; the Wilcoxon test remains significant at the 1% level for all metrics in both OOS splits, and the McNemar test is similarly significant at the 1% level across all contexts. Furthermore, the out-of-sample accuracy of set-2 (0.885–0.907) is remarkably high for a 28-class problem with only ten images per class. This demonstrates that the proposed framework provides a compact yet highly discriminative representation, especially notable given that the feature vector consists of only eight scalar values per image—a stark contrast to the thousands of coefficients required by conventional descriptors like Gabor filter banks or local binary pattern histograms.[53]

One of the definitive advantages of this proposed framework is its simplicity and low computational cost compared to the recent literature.[54,55] While hybrid approaches blending fractal features with neural networks have achieved higher peak accuracies, such as the 99% accuracy on the Kylberg dataset reported by Tiwari *et al.*,[54] such performance often comes at a significant practical cost. Unlike Convolutional Neural Networks (CNNs), which require large amounts of labeled data to train properly and involve intensive supervised training, our method utilizes a limited dataset combined with robust, interpretable features. This design allows for computationally efficient classification while mitigating the risk of overfitting inherent in small-dataset deep learning.

Several limitations of the present study should be acknowledged. First, although the reduction of each 512×512 image to eight scalar values prior to classification yields a highly interpretable feature vector grounded in information theory, it necessarily discards spatial heterogeneity within the image. Textures that are globally homogeneous but locally variable (or that contain distinct subregions with different statistical properties) cannot be distinguished by any global descriptor, regardless of its theoretical properties.

Second, although the Hilbert curve minimizes directional bias compared to row-wise scans, it still imposes a fixed traversal order on the image, and different space-filling curves may encode complementary aspects of the spatial correlation structure. Comparing ordinal pattern distributions obtained from multiple space-filling curves and treating them as independent feature streams is a direction that has received little attention in the literature and could yield meaningful gains.

Third, the Kylberg dataset consists of carefully controlled photographic images acquired under uniform illumination; applying the methodology to textures extracted from remote sensing imagery, histological slides, or industrial inspection systems, where acquisition conditions vary, and class boundaries are less clearly defined, would constitute a rigorous test of real-world applicability.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their insightful comments, which significantly enhanced the manuscript. This work was partially supported by Universidad de Buenos Aires under Grant No. UBACyT 20020220200045BA and by Universidad Nacional de La Plata under Project No. I286.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Ma. Belén Arouxet: Formal analysis (equal); Investigation (equal); Methodology (equal); Writing – original draft (equal); Writing – review & editing (equal). **Aurelio F. F. Bariviera:** Formal analysis (equal); Methodology (equal); Software (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Roberta Hansen:** Formal analysis (equal); Methodology (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Verónica E. Pastor:** Formal analysis

22 October 2026 09:21:48

© Author(s) 2026

(equal); Methodology (equal); Software (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

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

REFERENCES

[1] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **13**, 6308–6325 (2020).

[2] R. Lobos, J. Silva, J. Ortiz, G. Díaz, and A. Egaña, "Analysis and classification of natural rock textures based on new transform-based features," Math. Geosci. **48**, 835–870 (2016).

[3] E. T. C. Chagas, A. C. Frery, O. A. Rosso, and H. S. Ramos, "Analysis and classification of SAR textures using information theory," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **14**, 663–675 (2021).

[4] A. A. Pessa, R. S. Zola, M. Perc, and H. V. Ribeiro, "Determining liquid crystal properties with ordinal networks and machine learning," Chaos Soliton. Fract. **154**, 111607 (2022).

[5] A. T. Bui and D. W. Apley, "A monitoring and diagnostic approach for stochastic textured surfaces," Technometrics **60**, 1–13 (2018).

[6] C. Bandt and B. Pompe, "Permutation entropy: A natural complexity measure for time series," Phys. Rev. Lett. **88**, 174102 (2002).

[7] C. Bandt and K. Wittfeld, "Two new parameters for the ordinal analysis of images," Chaos **33**, 043124 (2023).

[8] H. V. Ribeiro, L. Zunino, E. K. Lenzi, P. A. Santoro, and R. S. Mendes, "Complexity-entropy causality plane as a complexity measure for two-dimensional patterns," PLoS One **7**, e40689 (2012).

[9] L. Zunino and H. V. Ribeiro, "Discriminating image textures with the multiscale two-dimensional complexity-entropy causality plane," Chaos Soliton. Fract. **91**, 679–688 (2016).

[10] A. F. Bariviera, R. Hansen, and V. E. Pastor, "Texture discrimination via Hilbert curve path based information quantifiers," Pattern Anal. Appl. **28**, 31 (2025).

[11] C. H. Weiß and P. Adämmer, "Nonparametric testing of spatial dependence in 2D and 3D random fields," Chaos **36**, 043116 (2026).

[12] B. Fadlallah, B. Chen, A. Keil, and J. Príncipe, "Weighted-permutation entropy: A complexity measure for time series incorporating amplitude information," Phys. Rev. E—Stat. Nonlinear Soft Matter Phys. **87**, 022911 (2013).

[13] H. Azami and J. Escudero, "Amplitude-aware permutation entropy: Illustration in spike detection and signal segmentation," Comput. Methods Programs Biomed. **128**, 40–51 (2016).

[14] D. Cuesta–Frau, "Permutation entropy: Influence of amplitude information on time series classification performance," Math. Biosci. Eng. **16**, 6842–6857 (2019).

[15] A. F. Bariviera, M. B. Guercio, L. B. Martinez, and O. A. Rosso, "A permutation information theory tour through different interest rate maturities: The Libor case," Philos. Trans. R. Soc. Lond. A **373**, 20150119 (2015).

[16] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, edited by A. J. Smola, P. J. Bartlett, B. Schölkopf, and D. Schuurmans (MIT Press, Cambridge, MA, 1999), pp. 61–74.

[17] V. Vapnik, S. E. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation and signal processing," in *Proceedings of the 10th International Conference on Neural Information Processing Systems, NIPS'96* (MIT Press, Cambridge, MA, 1996), pp. 281–287.

[18] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," Biochim. Biophys. Acta (BBA) - Protein Struct. **405**, 442–451 (1975).

[19] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika **12**, 153–157 (1947).

[20] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Comput. **10**, 1895–1923 (1998).

[21] H. Chen, I. C. Covert, S. M. Lundberg, and S.-I. Lee, "Algorithms to estimate Shapley value feature attributions," Nat. Mach. Intell. **5**, 590–601 (2023).

[22] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS'17, Long Beach, CA* (Curran Associates Inc., Red Hook, NY, 2017), pp. 4768–4777.

[23] C. F. B. Moon, H. V. Jagadish, and J. H. Saltz, "Analysis of the clustering properties of the Hilbert space-filling curve," IEEE Trans. Knowl. Data Eng. **13**, 124–141 (2001).

[24] H. Dai and H. Su, "Studies of norm-based locality measures of two-dimensional Hilbert curves," SN Comput. Sci. **2**, 403 (2021).

[25] Y. Rong, X. Zhang, and J. Lin, "Modified Hilbert curve for rectangles and cuboids and its application in entropy coding for image and video compression," Entropy **23**, 836 (2021).

[26] J. Červený, see https://github.com/jakubcerveny/gilbert for "Gilbert: Space-filling curve for rectangular domains or arbitrary size" (2025) (accessed June 09, 2026).

[27] A. M. Y. R. de Sousa and J. Hlinka, "Assessing serial dependence in ordinal patterns processes using chi-squared tests with application to EEG data analysis," Chaos **32**, 073126 (2022).

[28] A. Silbernagela and C. H. Weiß, "The joint asymptotic distribution of entropy and complexity," Chaos **36**, 023145 (2026).

[29] K. Schwarz, see http://www.keithschwarz.com/interesting/code/factoradic-permutation/FactoradicPermutation for "Factoradic permutation" (2024) (accessed 15/02/2024).

[30] F. Olivares, A. Plastino, and O. A. Rosso, "Ambiguities in Bandt–Pompe's methodology for local entropic quantifiers," Physica A **391**, 2518–2526 (2012).

[31] F. Olivares, A. Plastino, and O. A. Rosso, "Contrasting chaos with noise via local versus global information quantifiers," Phys. Lett. A **376**, 1577–1583 (2012).

[32] E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Champaign, IL, 1949).

[33] D. P. Feldman and J. P. Crutchfield, "Measures of statistical complexity: Why?" Phys. Lett. A **238**, 244–252 (1998).

[34] R. López-Ruiz, H. L. Mancini, and X. Calbet, "A statistical measure of complexity," Phys. Lett. A **209**, 321–326 (1995).

[35] P. Lamberti, M. Martin, A. Plastino, and O. Rosso, "Intensive entropic non-triviality measure," Physica A **334**, 119–131 (2004).

[36] M. Martín, A. Plastino, and O. Rosso, "Statistical complexity and disequilibrium," Phys. Lett. A **311**, 126–132 (2003).

[37] P. W. Lamberti, A. P. Majtey, A. Borras, M. Casas, and A. Plastino, "Metric character of the quantum Jensen-Shannon divergence," Phys. Rev. A **77**, 052311 (2008).

[38] O. A. Rosso, H. A. Larrondo, M. T. Martín, A. Plastino, and M. A. Fuentes, "Distinguishing noise from chaos," Phys. Rev. Lett. **99**, 154102 (2007).

[39] L. Zunino, M. C. Soriano, I. Fischer, O. A. Rosso, and C. R. Mirasso, "Permutation-information-theory approach to unveil delay dynamics from time-series analysis," Phys. Rev. E **82**, 046212 (2010).

[40] M. C. Soriano, L. Zunino, O. A. Rosso, I. Fischer, and C. R. Mirasso, "Time scales of a chaotic semiconductor laser with optical feedback under the lens of a permutation information analysis," IEEE J. Quantum Electron. **47**, 252–261 (2011).

[41] L. Zunino, B. M. Tabak, F. Serinaldi, M. Zanin, D. G. Pérez, and O. A. Rosso, "Commodity predictability analysis with a permutation information theory approach," Physica A **390**, 876–890 (2011).

[42] A. F. Bariviera, M. B. Guercio, L. B. Martinez, and O. A. Rosso, "The (in)visible hand in the Libor market: An information theory approach," Eur. Phys. J. B **88**, 208 (2015).

[43] R. A. Fisher, "On the mathematical foundations of theoretical statistics," Philos. Trans. R. Soc. Lond. A **222**, 309–368 (1922).

[44] B. R. Frieden, *Physics from Fisher Information: A Unification* (Cambridge University Press, Cambridge, 1998).

[45] P. Sánchez-Moreno, R. J. Yanez, and J. S. Dehesa, "Discrete densities and Fisher information," in *Proceedings of the 14th International Conference on Difference Equations and Applications*, edited by M. Bohner, Z. Dosla, G. Ladas, M. Ünal, and A. Zafer (Bahcesehir University Press, 2009), pp. 291–298.

22 October 2026 09:21:48

© Author(s) 2026

[46]C. Vignat and J.-F. Bercher, "Analysis of signals in the Fisher–Shannon information plane," Phys. Lett. A **312**, 27–33 (2003).
[47]C. Villani, *Topics in Optimal Transportation*, Graduate Studies in Mathematics (American Mathematical Society, Providence, RI, 2003), Vol. 58.
[48]S. Peleg, M. Werman, and H. Rom, "A unified approach to the change of resolution: Space and gray-level," IEEE Trans. Pattern Anal. Mach. Intell. **11**, 739–742 (1989).
[49]H. Yu, X. Huang, X. Hu, and H. Cai, "A comparative study on data mining algorithms for individual credit risk evaluation," in *2010 Fourth International Conference on Management of E-Commerce and E-Government (ICMeCG 2010)* (IEEE Computer Society, Los Alamitos, CA, USA, 2010), pp. 35–38.
[50]G. Kylberg, "The Kylberg texture dataset v. 1.0," Tech. Rep. 35 (Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, Uppsala, Sweden, 2011), see http://www.cb.uu.se/gustaf/texture/ and https://kylberg.org/datasets/.
[51]J. Gorodkin, "Comparing two *K*-category assignments by a *K*-category correlation coefficient," Comput. Biol. Chem. **28**, 367–374 (2004).
[52]D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics **21**, 1–13 (2020).
[53]G. Kylberg and I. M. Sintorn, "Evaluation of noise robustness for local binary pattern descriptors in texture classification," EURASIP J. Image Video Process. **2013**, 17 (2013).
[54]S. Tiwari, A. K. Sharma, I. Abdul Aziz, D. Gupta, A. Jain, H. Mahdin, S. Athithan, and R. Hidayat, "Investigations on segmentation-based fractal texture for texture classification in the presence of Gaussian noise," PLoS One **20**, e0315135 (2025).
[55]V. Gupta, A. Mishra, and N. Shrivastava, "Hytexnet: Percentile-guided local encoding and deep feature fusion for enhanced texture classification," Knowl.-Based Syst. **338**, 115482 (2026).

22 October 2026 09:21:48

© Author(s) 2026