ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Information-Theoretic Unsupervised Feature Selection for High-Dimensional Spatial Data

Samuel Suárez-Marcote[1], Abhijeet Vishwasrao[2], Ricardo Vinuesa[2], Laura Morán-Fernández[1] and Verónica Bolón-Canedo[1] *

1- CITIC, Univ. da Coruña, A Coruña, Campus de Elviña, 15071, Spain
2- Dept. of Aerospace Engineering, Univ. of Michigan, Ann Arbor, MI 48109, USA

Abstract. High-dimensional unlabelled datasets present significant challenges for efficient analysis, storage and interpretation. Unsupervised feature selection offers a way to retain the most informative variables while discarding redundant or uninformative ones, enabling more scalable processing. We introduce a spatially aware, unsupervised method that uses information theoretic criteria to identify informative variables while limiting redundancy, producing compact and spatially dispersed subsets of features. Our approach avoids dependence on labelled data or model-specific wrappers, making it suitable for large unstructured datasets. Experiments on MNIST and EMNIST datasets, including high-resolution upscaled versions, show that the selected features preserve both discriminative structure and reconstruction quality better than chosen supervised and unsupervised baselines, demonstrating the effectiveness of entropy and mutual information coupling in unlabelled high-dimensional settings.

1 Introduction

In recent years, advances in data collection technology and the fast growth of Internet of Things devices have led to a significant increase in the volume of raw data. While valuable, this growth introduces significant practical challenges, including prohibitive storage costs, increased processing latency and amplified effects of measurement noise [1]. A further complication is that much of this data is collected without labels, making supervised learning approaches impractical, as the labelling process is often time-consuming and expensive. To mitigate these limitations, dimensionality reduction methods are required. Specifically, feature selection has become a central tool, allowing informative features to be preserved while removing irrelevant or redundant ones, reducing computational costs and improving model explainability, thereby enabling scalable analysis when manual annotation is unavailable.

Feature selection is valuable in a wide range of real-world application domains. For example, in remote sensing and medical imaging, feature selection aids in identifying informative pixels or regions, enabling accurate analysis while

*This work was supported by Xunta de Galicia/FEDER (ED431C 2022/44); Ministerio de Ciencia e Innovación MICIU/AEI/10.13039/501100011033 and "Next-GenerationEU"/PRTR via Grant PID2023-147404OB-I00, Ministry for Digital Transformation and Civil Service (TSI-100925-2023-1) and Programa de axudas á etapa predoutoral, Xunta de Galicia (ED481A 2023/034). CITIC, as an accredited Galician University System Research Center, is supported by ERDF Funds and by "Secretaría Xeral de Universidades" (Grant ED431G 2023/01).

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

limiting data volume [2]. In environmental and urban monitoring, selecting a reduced yet meaningful set of sensor locations is essential for studying wind patterns, pollution dispersion, and urban heat islands. Such information is critical for sustainable urban planning and improving public health [3]. These applications highlight the relevance of feature selection not only for improving analytical performance but also for promoting resource efficiency. Reducing the number of measurements or focusing on the most informative locations decreases storage and computational requirements, minimises energy consumption and contributes to greener and more scalable data-driven systems while enhancing the robustness of downstream models.

With this context in mind, we propose a new unsupervised feature selection method designed for spatially structured, large-scale datasets. As a filter method, the approach does not depend on any particular induction algorithm, which reduces computational cost by decoupling selection from model training. Common supervised strategies typically do not exploit the spatial structure of image data, producing dense and spatially clustered feature sets that are redundant for reconstruction and analysis. Our method consists of three simple stages that prioritise scalability and minimise redundancy. First, an information-based score is computed at each spatial location to identify promising informative pixels without relying on labels. Second, spatial coherence is exploited by evaluating overlapping local patches and retaining a set of high-information candidate regions. This step significantly reduces the search space and the amount of subsequent computation needed. Third, selection among candidates balances relevance with pairwise dependence to avoid choosing multiple regions that provide the same information. By operating as an unsupervised filter that combines entropy driven screening with mutual information based redundancy control, the proposed method produces a small, interpretable, and spatially dispersed set of locations while remaining computationally practical for large scale datasets.

2 Proposed method

In unsupervised feature selection, the goal is to uncover patterns and structures from correlations in unlabelled data. Common statistical criteria include variance, entropy and mutual information [4]. Our framework specifically examines the synergy between entropy and mutual information. While mutual informationoffers an explicit measure of shared information and redundancy among candidate features, entropy quantifies individual uncertainty and relevance. Unlike traditional methods that treat features as independent vectors, our approach exploits the spatial coherence inherent in image grids and sensor arrays.

The first stage evaluates the raw information content at each spatial location. For every feature with coordinates (x, y) we compute the Shannon entropy across the N samples in the dataset. To ensure scalability for large high-resolution datasets, we utilise a memory efficient estimator that processes samples iteratively. This results in a global entropy map, $H(x, y)$, which highlights regions of high statistical activity while filtering out static background noise.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

To mitigate the effects of pixel-level noise and capture local structural dependencies while significantly reducing the number of necessary calculations, we aggregate pixels into overlapping patches. Let P denote a patch as the set of pixel indices within a local window anchored at coordinates (x, y). For each one, we associate a low-dimensional descriptor that summarises the patch behaviour across the sample set. The choice of patch descriptor affects the determination of which spatial locations are considered informative and which are seen as redundant. While simple one-dimensional summaries can capture essential patch structures, more expressive descriptors, such as PCA or embeddings, can better encapsulate detailed within-patch information while reducing sensitivity to sample noise when necessary. We denote the descriptor time series for patch P across the N samples as $d_P = (d_P^1, \ldots, d_P^N)$.

After screening, candidate patches are selected iteratively under a trade-off between relevance and redundancy [5]. Relevance of a candidate patch P is measured by the patch descriptor entropy $H(d_P)$ while redundancy between two candidate patches P and Q is measured by their mutual information:

$$I(d_P; d_Q) = \sum_u \sum_v p(u, v) \log \frac{p(u, v)}{p(u)p(v)} \tag{1}$$

where $p(u, v)$ is the joint probability that the descriptor of patch P takes value u and the descriptor of patch Q takes value v and where $p(u)$ and $p(v)$ are the corresponding marginal probabilities.

A minimum Redundancy Maximum Relevance selection criterion chooses at each step the candidate, P, that maximises a score of the form:

$$\text{score}(P) = H(d_P) - \max_{Q \in S} I(d_P; d_Q), \tag{2}$$

where S is the set of already selected candidates. The subtraction of the maximum pairwise mutual information penalises candidates that are redundant with any already selected region, encouraging a diverse set of informative patches.

Finally, once patches are selected, representative pixels are chosen by returning to the pixel-level entropy map and selecting locations of maximal $H(x, y)$ within each patch. This step ties the patch-level decision back to the original uncertainty measure and helps ensure spatial dispersion.

3 Experimental evaluation

This section describes the experimental process and a comprehensive analysis of the results. Our method is tested across several image resolutions and compared with both supervised and unsupervised feature selection methods.

3.1 Experimental settings

To evaluate the efficacy of the proposed method, we conducted a comparative analysis on two standard datasets. The MNIST dataset [6] contains 70,000 images across 10 classes, while EMNIST [7] provides a more heterogeneous scenario

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

with 131,600 samples distributed over 47 classes. To assess the scalability and robustness of the method in high-dimensional spaces, we evaluated the datasets at their original resolution of 28×28 and at upscaled resolutions of 48×48 and 64×64. These larger grids intentionally introduce additional irrelevant pixels and dilute individual feature information, creating a more challenging setting for feature selection and reconstruction. Although the proposed method is unsupervised, we evaluate it on labelled datasets. This allows objective measurements of how well the selected features preserve discriminative structure and enables fair comparison against supervised and unsupervised baselines.

For the sake of brevity, the number of selected features was fixed at 10% of the total available pixels for all experiments. This approach allowed for testing the identification of the most critical information with a reduced number of features. Performance was evaluated using three complementary metrics: (I) classification accuracy was chosen to quantify the discriminative information preserved by the selected pixels, measured by training a Random Forest classifier on the reduced feature space, (II) relative Mean Squared Error (rMSE) captured the reconstruction fidelity, reflecting how much of the original image can be recovered by a lightweight autoencoder when only the selected pixels are provided, and (III) reconstructed accuracy evaluated the preservation of class-relevant structure by applying the same classifier to the autoencoder-reconstructed images, thus indicating whether the selected features retain not only raw intensity information but also the spatial patterns necessary for recognition.

For reliability, each configuration is evaluated across a 3-repetition 5-fold cross-validation, which involves carrying out feature selection, classification and reconstruction steps within a single cross-validation loop. We compare our approach ("Ours") against minimum Redundancy Maximum Relevance (mRMR) [5], as a labelled feature selection baseline, and a simple variance ranking (Var.) [8], selecting the highest variance features, as a representative unlabelled method.

3.2 Results and analysis

Table 1 shows the mean and standard deviation for the mentioned metrics. On MNIST dataset, our method achieves the highest classification accuracy at every resolution, indicating that the selected pixels preserve more discriminative information. The rMSE is substantially lower than for mRMR and variance ranking, showing that the selected features retain a greater proportion of the original signal, which is further reflected in the superior reconstructed accuracy. These trends become even more pronounced as the input resolution increases, suggesting that the method scales effectively and remains robust in higher-dimensional settings. An illustrative example is given in Figure 1, which presents an original MNIST sample alongside the reconstructions produced by each method. As shown there, the pixels chosen by the proposed method are visibly more spatially dispersed and the reconstructed images contain less noise than the other methods, corroborating the quantitative results. A similar behaviour is observed on the EMNIST dataset. Although mRMR obtains slightly higher classification accuracy and lower rMSE at this resolution, our proposed approach maintains a

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

clear advantage at 48×48 and 64×64 resolutions. The reduction in reconstruction error and accuracy demonstrate that the method remains effective even in more challenging classification scenarios with higher intra-class variability. Overall, the results confirm that entropy-guided patch screening combined with redundancy control produces feature sets that are both more informative and less redundant than those produced by conventional unsupervised baselines.

Dataset	Size	Method	Classification Accuracy	rMSE	Reconstructed Accuracy
MNIST	28x28	Var.	0.916 ± 0.002	0.166 ± 0.004	0.662 ± 0.032
		mRMR	0.920 ± 0.002	0.159 ± 0.003	0.719 ± 0.040
		Ours	**0.944** ± 0.002	**0.114** ± 0.003	**0.735** ± 0.034
	48x48	Var.	0.939 ± 0.001	0.098 ± 0.001	0.694 ± 0.034
		mRMR	0.930 ± 0.001	0.110 ± 0.001	0.739 ± 0.033
		Ours	**0.951** ± 0.001	**0.045** ± 0.002	**0.796** ± 0.026
	64x64	Var.	0.940 ± 0.001	0.133 ± 0.001	0.705 ± 0.033
		mRMR	0.934 ± 0.001	0.135 ± 0.002	0.706 ± 0.034
		Ours	**0.958** ± 0.001	**0.044** ± 0.003	**0.800** ± 0.026
EMNIST	28x28	Var.	0.575 ± 0.003	0.236 ± 0.004	0.344 ± 0.003
		mRMR	**0.725** ± 0.005	**0.121** ± 0.004	**0.437** ± 0.006
		Ours	0.706 ± 0.003	0.142 ± 0.002	0.414 ± 0.001
	48x48	Var.	0.624 ± 0.011	0.168 ± 0.007	0.366 ± 0.019
		mRMR	0.723 ± 0.004	0.074 ± 0.010	0.434 ± 0.017
		Ours	**0.759** ± 0.001	**0.029** ± 0.004	**0.438** ± 0.012
	64x64	Var.	0.830 ± 0.007	0.068 ± 0.004	0.731 ± 0.011
		mRMR	0.831 ± 0.004	0.068 ± 0.004	0.738 ± 0.009
		Ours	**0.857** ± 0.001	**0.027** ± 0.003	**0.750** ± 0.010

Table 1: Mean and standard deviation of performance metrics for the proposed method compared to mRMR and Var. baselines across MNIST and EMNIST datasets at 28×28, 48×48 and 64×64 resolutions. The number of selected features is fixed at 10% of the total available pixels. Best results shown in bold.

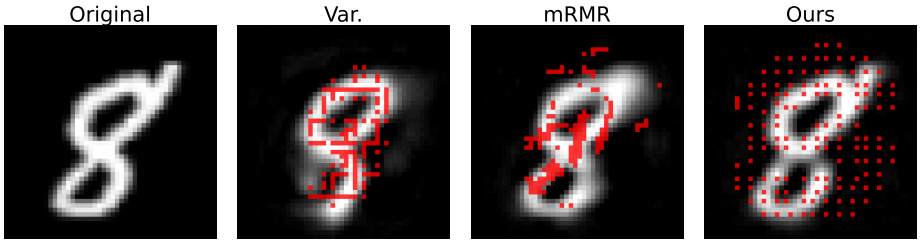

Fig. 1: Visual comparison of an MNIST sample and its reconstruction using the proposed method (Ours) against supervised (mRMR) and unsupervised (Var.) baselines. Features selected by each algorithm are highlighted in red.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

4 Conclusions

We introduced a scalable unsupervised spatial feature selection method that resorts to information-theoretic criteria to identify and retain informative variables while limiting redundancy. The method produces compact and spatially dispersed sets of informative features without reliance on labels or heavy model-specific wrappers. Experiments on MNIST and EMNIST datasets, including their upscaled variants, show improved preservation of discriminative structure and better reconstruction fidelity than common supervised and unsupervised baselines, with advantages that increase at higher resolution.

Although the study is preliminary, the approach has clear practical benefits for monitoring systems. Selecting a small set of high-information locations reduces data acquisition, communication and storage demands, and thus the energy footprint of sensing infrastructures. This makes the method attractive for sensor-placement applications in urban environments, for example, airflow and pollution monitoring, where a limited number of well-placed sensors can substantially reduce deployment and operational costs while still capturing the dominant spatial patterns. As a first future step, we plan to apply the method to urban airflow data to evaluate its effectiveness in complex and domain-specific scenarios. Additionally, other directions remain open for further development. These include adapting the framework to multichannel data, exploring richer descriptors such as multi-scale or learned representations and integrating domain-specific constraints like energy consumption and deployment cost.

References

[1] Iain M. Johnstone and D. Michael Titterington. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A*, 367:4237–4253, 2009.

[2] Jie Feng, Licheng Jiao, Fang Liu, Tao Sun, and Xiangrong Zhang. Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images. *Pattern Recognition*, 51:295–309, 2016.

[3] Pablo Torres, Soledad Le Clainche, and Ricardo Vinuesa. On the experimental, numerical and data-driven methods to study urban flows. *Energies*, 14(5):1310, 2021.

[4] Guojie Li, Zhiwen Yu, Kaixiang Yang, Mianfen Lin, and CL Philip Chen. Exploring feature selection with limited labels: A comprehensive survey of semi-supervised and unsupervised approaches. *IEEE Trans. on Knowledge and Data Engineering*, 36(11):6124–6144, 2024.

[5] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[6] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[7] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017.

[8] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.