Rethinking HyperSpectral Image Classification (HSIC) Benchmark with Explainability (xAI) under a Causal Estimation Perspective

Sékou Dabo[1], Michele Linardi[1] and Claudia Paris[2]

[1]ETIS, UMR 8051, CYU, ENSEA, CNRS, Cergy, France
[2]University of Twente, ITC, 7522NH Enschede, The Netherlands

Abstract

In this paper, we revisited patch-based hyperspectral image classification from a causal perspective, grounded in the intrinsic assumption that the central pixel and its local neighborhood constitute the main source of the predicted label. By explicitly controlling spatial leakage through a disjoint data partitioning protocol, we showed that common evaluation practices not only inflate performance metrics but also alter the causal structure learned by state-of-the-art models.

Our causal sensitivity analysis reveals that, in the absence of leakage, models consistently rely on the Region of Interest to support their predictions, whereas under random sampling this causal link becomes obscured by confounding spatial correlations.

These findings highlight that reliable assessment of HSI classifiers requires both leakage-free evaluation and causal scrutiny of learned representations. Beyond improving benchmarking practices, this work calls for a rethinking of patch-based HSI classification toward models and protocols that favor causal fidelity over shortcut-driven performance.

Keywords

Satellite Hyperspectral Imaging (HSI), Deep Learning (DL), Crop Type Mapping, Earth Observation (EO) data, Explainability (xAI)

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

1. Introduction

Hyperspectral imaging (HSI) analysis plays an increasingly important role in domains such as precision agriculture, environmental monitoring, urban mapping, and medical diagnosis [1]. The integration of deep learning has enabled a shift beyond purely spectral feature analysis (*pixel-wise* classification), toward approaches that jointly exploit both spectral and spatial information. Deep vision models (2D/3D CNNs, hybrid 1D+2D architectures, Vision Transformers and their variants) have led to the widespread adoption of the *patch-wise* paradigm, where the label of a central pixel is inferred from its surrounding spatial context.

However, due to the high cost of annotation and the scarcity of acquisition campaigns, datasets available for training and evaluation often consist of a single scene. Consequently, model evaluation is highly dependent on how this scene is divided into training and testing subsets.

Pixel-wise random sampling, yet commonly used in the literature, introduces train-test leakage, caused by the spatial proximity of samples, which results in overestimated performance [2, 3, 4, 5, 6].

To address this issue, several works have proposed *disjoint object-level splitting* strategies [7, 8, 9, 10, 5, 4, 3] that spatially separate training and testing regions.

Published in the Proceedings of the Workshops of the EDBT/ICDT 2026 Joint Conference (March 24-27, 2026), Tampere, Finland
dabo.sekou@cyu.fr (S. Dabo); michele.linardi@cyu.fr (M. Linardi); c.paris@utwente.nl (C. Paris)

Such methods aim to mitigate three types of leakage: (1) spatial proximity, (2) patch context, and (3) structural leakage arising from spatial autocorrelation. Addressing all three simultaneously remains an open challenge.

Block-based or cluster-based data partitioning sampling reduce spatial dependence by enforcing geographical separation between training and test samples. However, they may still allow boundary-induced patch overlap, leading to contextual leakage, and can result in the exclusion of certain classes from either the training or test sets when class distributions are spatially unbalanced [7, 5]. In contrast, controlled or continuous sampling strategies applied over the entire scene distribute boundary effects across the image, but do not eliminate spatial autocorrelation [11, 6]. As a consequence, strong dependencies between neighboring samples persist, which can substantially underestimate the true generalization error [3, 4].

We argue that the presence of train-test leakage accentuates the *opaque* modelling of deep neural networks, which integrate a complex decision-making process involving high-dimensional non-linear interactions that remain difficult to interpret, even when the model achieves high accuracy.

In this sense, the field of Explainable AI (xAI) has gained significant attention across computer vision, natural language processing, and machine learning in general [12, 13, 14, 15].

However, several studies [16, 17], have shown that models often rely on shorcuts, namely spurious correlations present in the training data rather than genuine task-relevant features.

Such shortcuts can arise from (i) poor data preprocessing leading to annotation artefacts [17], or (ii) selection biases, such as non-representative or discriminative sampling strategies [16]. While such correlations can artificially inflate performance under a given data distribution, they frequently fail to generalize in real-world deployment scenarios.

In our work, we study the current state-of-the-art patch-based HSI classifiers and how to transparently explain the sensitivity to relevant data features. Specifically, we consider Landcover and Crop Mapping, namely critical tasks that supports precision agriculture, yield estimation, and policy decisions. In Figure 1, we depict an exemple of land cover classification on University of Pavia dataset, namely a urban hyperspectral benchmark acquired by the ROSIS sensor over the campus area in Pavia, northern Italy [18]. By design, deep learning classifiers are sensitive to biases introduced at dataset construction. As a result, the models are prone to exploiting undesired shortcuts that cause overfitting due to the limited number of unique examples, learning spurious patterns rather than the spectral–spatial characteristics strictly associated with target pixels. A known problem in this sense is the Hughes phenomenon [19, 6], where highly overparameterized models may overfit to nearly identical samples across training and testing splits.

In this paper, we make the following contributions:

- **Leakage-free evaluation protocol for reliable HSI classification.** We introduce a leakage-free data partitioning protocol for hyperspectral image classification that enforces strict pixel-level independence between training and testing sets. By eliminating boundary-induced patch overlap and long-range spatial dependencies, the proposed protocol enables a controlled and unbiased assessment of model performance and isolates the true impact of data leakage.
- **Causal framework to quantify leakage-induced shifts in model behavior.** Building on the intrinsic design of patch-based classification where the central pixel and its local neighborhood (Region of Interest, ROI) define the prediction target we formalize the expected causal structure in which ROI features should drive the output. We then perform an intervention-based causal sensitivity analysis using the Average Model Intervention Effect (AMIE) to measure how this causal relationship changes under leakage-prone versus leakage-free evaluation settings, quantifying the model's reliance on task-relevant versus spurious features.
- **xAI-driven discovery of systematic spatial bias patterns.** Finally, leveraging feature-attribution explainability (xAI), we uncover consistent and interpretable spatial patterns induced by data leakage across three benchmark datasets. Our analysis reveals how leakage distorts learned representations by redirecting attention beyond the ROI, providing qualitative evidence of hidden spatial biases in model decision-making.

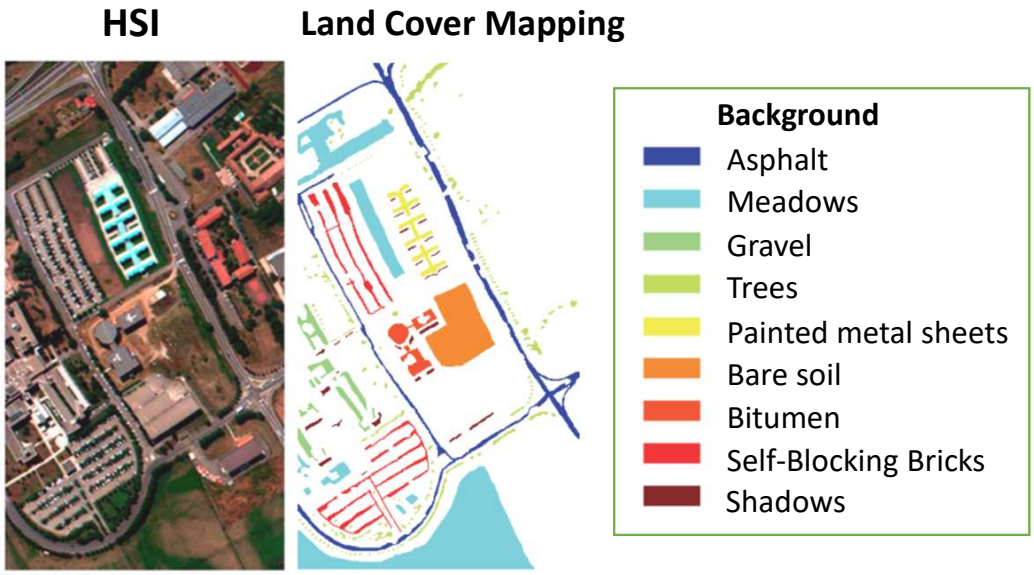

Figure 1: Land Cover mapping on the University of Pavia Dataset

2. Related Work

From Pixel-wise to Patch-based HSI Classification Hyperspectral image (HSI) classification has undergone a significant evolution, shifting from pixel-wise to patch-based paradigms in order to better exploit spatial context. In early pixel-wise approaches, each pixel is classified independently using its spectral vector as input to conventional machine learning models such as Support Vector Machines (SVMs) or 1D CNN / RNN-based architectures [20, 21, 22]. While effective in modeling spectral information, these methods inherently ignore spatial dependencies between neighboring pixels, limiting their ability to capture local structures.

To overcome this limitation, patch-based classification was introduced. In this paradigm, a spatial neighborhood (patch) is constructed around each central pixel and used as model input, while the label remains assigned solely to the central pixel [23, 24, 25, 26, 27, 28, 29]. This formulation enables models to jointly exploit spectral and spatial information and has become the dominant setting in modern HSI classification. However, it also introduces new challenges related to patch construction and, critically, to data sampling strategies.

Information Leakage Induced by Patch-based modeling and uniform random pixelwise Sampling The patch-based classification paradigm combined with random sampling can induce severe information leakage: spatially adjacent samples may share highly similar or overlapping patches, violating the independence assumption between training and test sets.

Several studies have shown that such leakage leads to a systematic overestimation of model performance, with reported accuracy drops of up to 10% once leakage is controlled [2, 3, 4, 5, 6]. These findings have motivated the development of spatially-aware sampling strategies aimed at enforcing a more realistic separation between training and testing data.

Spatially Disjoint Sampling Strategies Numerous sampling strategies have been proposed to mitigate spatial leakage while preserving data diversity. These include global spatial partitioning methods [7, 5] that enforce deterministic train–test separation with buffer zones, region-growing approaches that aim to limit overlap while maintaining spectral variability, and cluster-based strategies that reduce local spa-

tial dependence through class-wise grouping. Other methods rely on contiguous regional sampling [4, 3] to increase variability while partially reducing leakage. Despite their differences, these approaches share common limitations, such as sensitivity to patch size, constrained class balance, residual contextual overlap, and limited control over large-scale spatial autocorrelation. Collectively, these studies demonstrate that stricter spatial separation consistently yields lower—but more realistic—performance estimates, which better reflect the true generalization ability of modern models.

Limitations of Existing Evaluations and Role of XAI Despite these advances, most prior studies focus primarily on quantifying the performance drop induced by leakage removal. In parallel, explainable AI (xAI) methods have been widely applied in remote sensing to analyze spectral patterns of materials or to highlight salient spatial–spectral regions influencing model predictions [30, 31]. However, qualitative analyses of how information leakage alters the internal decision mechanisms of HSI classifiers remain largely unexplored.

In particular, existing works do not leverage XAI or causal analysis to examine whether leakage encourages models to rely on spurious spatial shortcuts, nor how leakage-free protocols affect the spatial fidelity of learned representations. Addressing this gap is essential for understanding not only how well models perform, but why they perform as they do under different evaluation protocols.

3. Methodology

Let $\mathcal{I} \in \mathbb{R}^{B\times H\times W}$ denote a hyperspectral image (HSI) composed of B spectral bands and a single observed scene of $H \times W$ pixels. Each spatial location (u, v) is associated with a spectral vector $\mathbf{x}_{u,v} \in \mathbb{R}^B$ and a ground-truth class label $y_{u,v} \in \mathcal{C} : \{c_i, \ldots, c_k\}$, where c_k is the k^{th} land-cover or crop class. Following the patch-based classification paradigm, a square spatial neighborhood (patch) $\mathbf{X}^{(s)}_{u,v} \in \mathbb{R}^{B\times S\times S}$ of size $S \times S$, where S must be odd and extracted around each central pixel (u, v) and used as input to a classifier, while the label remains that of the central pixel. Let $f_\theta : \mathbb{R}^{B\times S\times S} \to \mathcal{Y}$ denote a Machine Learning model parameterized by θ, producing class posterior probabilities $\mathcal{Y}$ over the set $\mathcal{C}$, with $\sum_{y_c\in\mathcal{Y}} y_c = 1$.

The learning problem considered in this work is land/crop mapping from HSI data under the realistic yet challenging setting where both training and test samples originate from a single scene. In this context, model evaluation is highly sensitive to the data partitioning strategy, as spatial dependencies between samples can introduce systematic biases in performance estimation. In line with prior work [6], we adopt a formal framework to characterize biases induced at the data-splitting stage and identify three main types of information leakage that we depict in Figure 2.

Type 1 – Spatial proximity leakage: training and test samples may lie sufficiently close in space such that their corresponding patches become nearly identical in terms of spatio-spectral content, effectively sharing the same information [6], see Figure 2(a).

Type 2 – Contextual leakage: even when spatially disjoint partitions are enforced, patches extracted from adjacent blocks belonging to different sets may still overlap at the pixel level, leading to unintended exposure of contextual information across training and test samples [6], see Figure 2(b).

Type 3 – Spatial autocorrelation: on the one hand, according to the first law of geography, nearby regions tend to exhibit similar spectral signatures; on the other hand, these regions often share comparable crop-configuration patterns. Consequently, regardless of whether random or disjoint splitting strategies are employed, samples drawn from nearby locations remain highly correlated, even in the absence of shared pixels, particularly when sampling is performed over the entire hyperspectral scenes.

3.1. Splitting strategies

We propose a holistic strategy that addresses multiple leakage sources, overcoming the limitation of existing data partitioning approaches, which do not address leakage simultaneously depending on spatial proximity, contextual overlap, and long-range spatial autocorrelations.

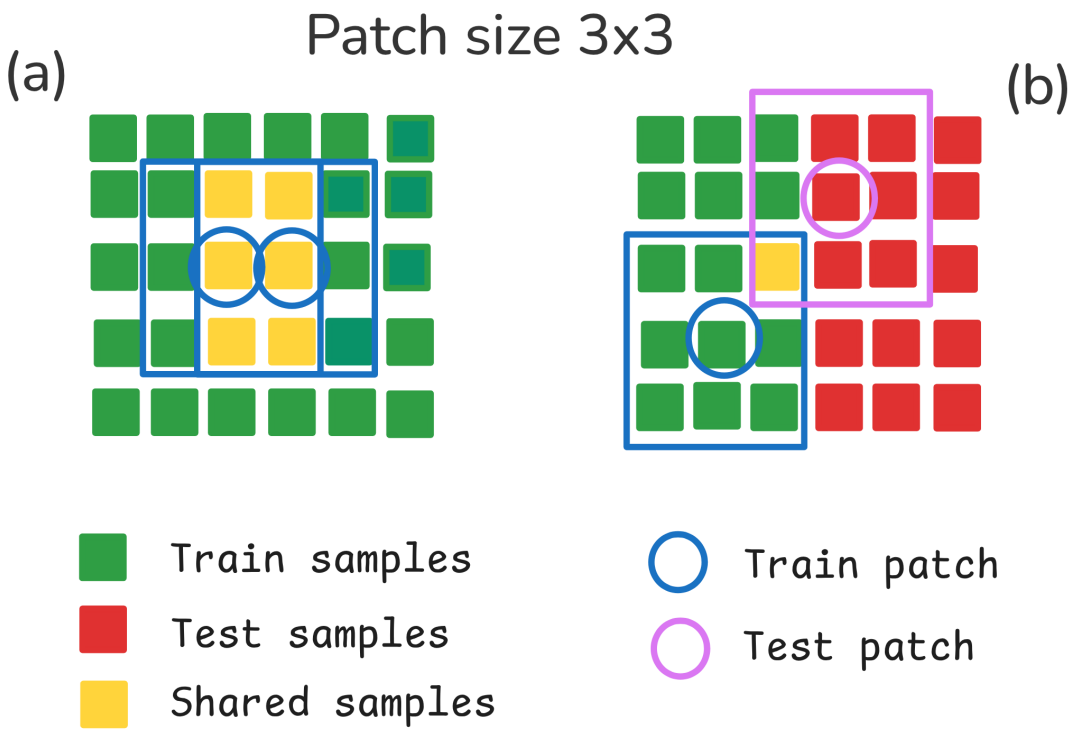

Figure 2: *(a) Type 1 Leakage (Spatial proximity) , (b) Type 2 Leakage (Contextual Leakage)*

Split3 relies on a spatially disjoint partitioning of the hyperspectral scene into two non-overlapping zones, one assigned to training and the other to testing. Instead of random sampling across the entire image, samples are progressively selected from one side of the scene toward the opposite side, ensuring that the test set corresponds to a spatial region never observed during training, preventing identical patches from appearing in both sets, while reducing dependencies induced by spatial autocorrelation.

The resulting setting resembles a localized domain generalization scenario, where the model must extrapolate to spatially distinct areas. To preserve class representativity without violating spatial disjointness, the selection process is conducted independently for each class.

Although zone-based partitioning mitigates spatial proximity and autocorrelation leakage, contextual overlap may still occur near region boundaries when constructing neighborhood-based patches. In this sense, we depict an example in Figure 3. To eliminate residual dependencies, boundary pixels within a margin equal to half the patch size are masked during patch extraction. This guarantees that no pixel contributes to both training and testing patches, enforcing strict pixel-level independence.

By jointly combining spatial disjoint partitioning and boundary masking, Split3 provides a controlled experimental protocol that simultaneously addresses all three leakage sources, enabling an unbiased evaluation and the isolation of leakage-induced performance inflation.

3.2. Causal Task Formulation and Region of Interest

In this part, we present the proposed causal sensitivity analysis framework based on the interventional paradigm on ROI features.

With an abuse of notation, we can denote a full patch with $\mathbf{X}$. Subsequently, we define $\text{ROI}_i \subset \mathbf{X}$, namely a square region of size $(2i+1) \times (2i+1)$ centered on the target pixel, with ROI_0, corresponding to a single pixel region.

By construction, patch-based state-of-the-art HSI classifiers leverage spectral–spatial features to predict the label of a single target pixel, which is in general the center of the patch.

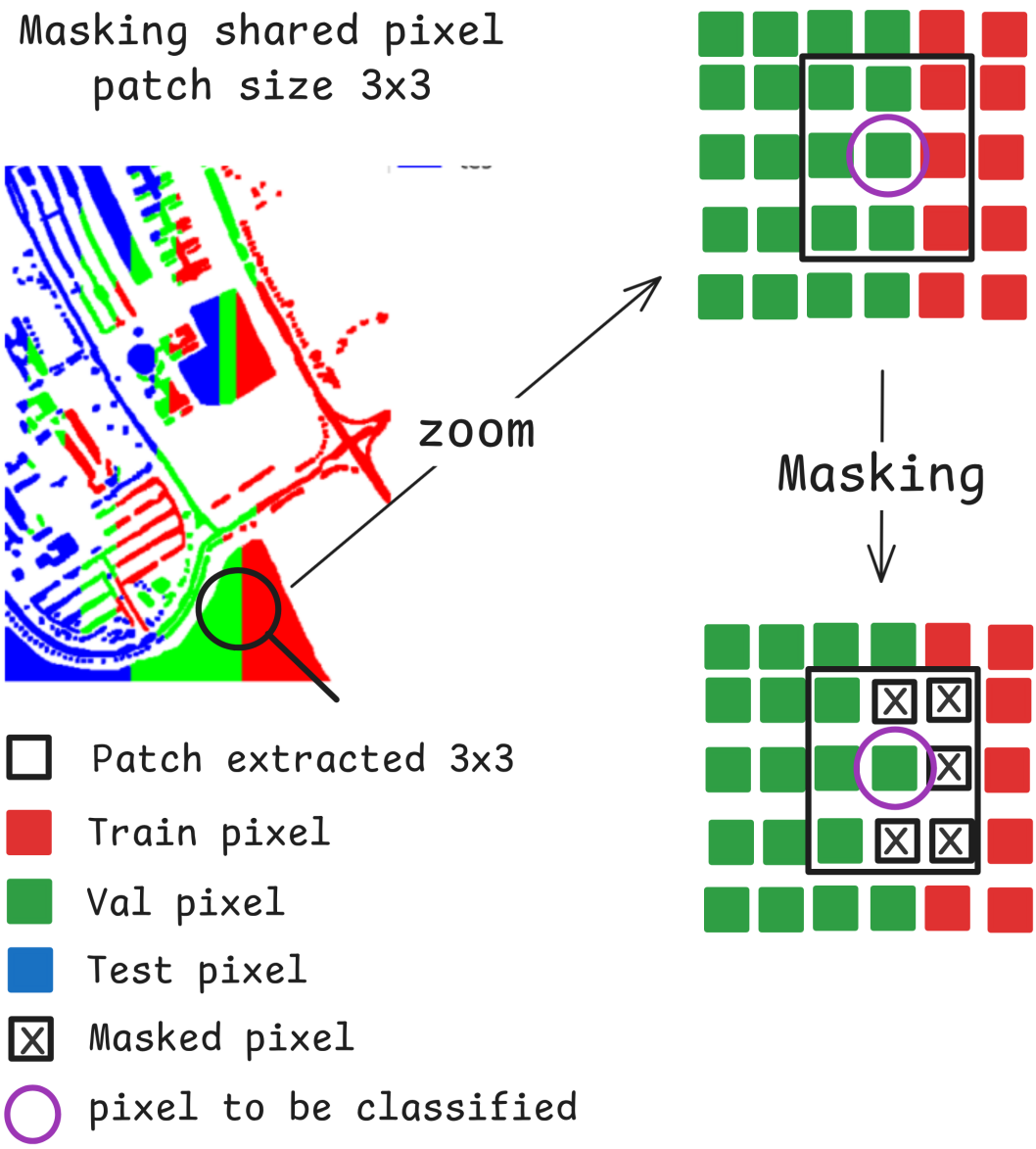

Figure 3: Split3 masking strategy to avoid boundary (train/test/val) pixels leakage. The figure depicts a masking occurring on val/train boundary.

Such a modelling design naturally induces an expected causal structure in which the ROI feature signature constitutes the cause of the predicted label, while the remaining pixels in the patch act as contextual variables.

Although contextual pixels may correlate with the label due to spatial organization and crop configuration patterns, they are not required to determine the class of the central pixel. Consequently, a model that relies predominantly on information outside the ROI can be considered as sensitive to spurious correlation rather than leveraging task-consistent causal pathways.

3.3. Causal Sensitivity to the Region of Interest

To quantify the extent to which a trained model adheres to the expected causal dependency between the ROI and the predicted label, we adopt an intervention-based analysis.

Hence, for a given patch $\mathbf{X}$, we define an intervention $T_{ROI_i}(\mathbf{X})$ that consists into altering the

information inside ROI_i replacing the ROI with a baseline value X_0, yielding the intervened input

$$T_{ROI_i}(\mathbf{X}) = \big(\mathbf{X} \setminus \mathrm{ROI}_i\big) \cup X_0(\mathrm{ROI}_i).$$

In our settings, we consider $X_0 = 0$ indicating the absence of material reflectance in the hyperspectral sensors.

Under this formulation, we use the Average Model Intervention Effect (AMIE) [32] score to measure the expected change in the model output induced by an intervention. Therefore we have:

$$\mathrm{AMIE}(T_{ROI_i}(\mathbf{X})) = \mathbb{E}_{\mathbf{X}}[p_\theta(y \mid \mathbf{X})] - \mathbb{E}_{\mathbf{X}}[p_\theta(y \mid T_i(\mathbf{X}))]\,,$$

where $y \in \mathcal{Y}$ denotes the ground-truth class of the central pixel. A large positive AMIE indicates that suppressing the ROI leads to a substantial decrease in the model's confidence for the correct class, reflecting a strong causal dependence on the ROI. Conversely, a low or negligible AMIE suggests that the model's predictions are largely driven by contextual information outside the ROI.

Within the proposed framework, AMIE does not serve to recover an unknown causal structure, but rather to evaluate whether and how the model's effective causal behavior deviates from the task-consistent relationship $\mathrm{ROI} \to Y$, particularly under different data partitioning strategies for model training.

3.4. Attribution Analysis

While AMIE quantifies the causal sensitivity of a model to the Region of Interest (ROI), it does not reveal how predictive information is spatially distributed within input patches. To complement the intervention-based analysis, we perform a feature attribution study to visualize spatial importance patterns.

Formally, for a given input patch $\mathbf{X}$, an attribution method $\mathcal{A}$ assigns to each pixel location (u, v) of $\mathbf{X}$ a relevance score indicating the contribution of that pixel to $f_\theta(X)$. We thus have :

$$\mathcal{A}\big(f_\theta(X)\big) \in \mathbb{R}^{S \times S}$$

Although attribution methods such as Integrated Gradients[33] provide instance-level (patch-wise) explanations, our objective is not to interpret individual predictions. Instead, we aim to characterize the model's global spatial behavior under different evaluation protocols. To this end, attribution maps are aggregated across test samples to estimate the expected spatial importance distribution of the model. Averaging attributions across instances (and across classes) provides a statistical descriptor of where the model systematically allocates importance within patches. In this sense, the aggregated heatmap in an empirical estimate of the expected attribution map under the data distribution.

The attribution analysis serves two purposes. First, it assesses whether the model's spatial focus aligns with the expected causal structure (dominant reliance on the ROI). Second, it provides a visual interpretation of AMIE trends. Models with strong causal dependence on the ROI are expected to produce centrally concentrated attribution maps, whereas diffuse or off-center patterns indicate reliance on contextual shortcuts.

Attribution maps are not interpreted as causal evidence in isolation but as complementary diagnostics supporting the intervention-based analysis.

4. Experimental Evaluation

In this section, we present the evaluation of state-of-the-art hyperspectral image (HSI) classification models under two data partitioning strategies: the commonly used *random sampling protocol*, which introduces spatial bias, and our proposal, namely Split3. Our empirical work aims to show the impact of the leakage types on the model's overall performance. Subsequently, by adopting our causal estimation framework based on the AMIE, we explain how leakage prevention affects feature reliance of the analyzed models under the lens of the ROI effect on the final output.

3pt

Table 1
Classification performance on **Indiana Pines (IP)**, **Salinas (SA)**, and **Pavia University (PU)** datasets under random and disjoint sampling protocols. Results are reported as mean ± standard deviation.

		DSFormer		SSRN		Hamida et al. (3D CNN)		ViT		SpectralFormer	
Dataset	Metric	Random	Split3	Random	Split3	Random	Split3	Random	Split3	Random	Split3
IP	OA	0.99 ± 0.00	0.45 ± 0.06	1.00 ± 0.00	0.65 ± 0.04	0.96 ± 0.00	0.53 ± 0.06	0.95 ± 0.00	0.45 ± 0.04	0.96 ± 0.00	0.39 ± 0.02
	AA	0.98 ± 0.00	0.33 ± 0.04	0.99 ± 0.00	0.54 ± 0.05	0.94 ± 0.00	0.43 ± 0.09	0.92 ± 0.02	0.29 ± 0.05	0.96 ± 0.01	0.20 ± 0.07
	F1 Macro	0.98 ± 0.00	0.31 ± 0.03	1.00 ± 0.00	0.54 ± 0.06	0.94 ± 0.00	0.42 ± 0.10	0.93 ± 0.01	0.27 ± 0.05	0.96 ± 0.01	0.17 ± 0.09
SA	OA	1.00 ± 0.00	0.85 ± 0.04	1.00 ± 0.00	0.83 ± 0.03	0.99 ± 0.00	0.86 ± 0.04	0.99 ± 0.00	0.82 ± 0.08	0.99 ± 0.00	0.83 ± 0.04
	AA	1.00 ± 0.00	0.91 ± 0.04	1.00 ± 0.00	0.90 ± 0.04	0.99 ± 0.00	0.92 ± 0.03	0.99 ± 0.00	0.87 ± 0.05	1.00 ± 0.00	0.88 ± 0.03
	F1 Macro	1.00 ± 0.00	0.90 ± 0.04	1.00 ± 0.00	0.89 ± 0.04	0.99 ± 0.00	0.91 ± 0.03	0.99 ± 0.00	0.86 ± 0.05	1.00 ± 0.00	0.87 ± 0.04
PU	OA	1.00 ± 0.00	0.64 ± 0.18	1.00 ± 0.00	0.70 ± 0.11	1.00 ± 0.00	0.67 ± 0.09	0.99 ± 0.00	0.63 ± 0.07	0.99 ± 0.00	0.65 ± 0.06
	AA	1.00 ± 0.00	0.74 ± 0.07	1.00 ± 0.00	0.81 ± 0.05	0.99 ± 0.00	0.78 ± 0.07	0.99 ± 0.00	0.76 ± 0.07	0.99 ± 0.00	0.77 ± 0.06
	F1 Macro	1.00 ± 0.00	0.66 ± 0.09	1.00 ± 0.00	0.74 ± 0.08	0.99 ± 0.00	0.71 ± 0.08	0.99 ± 0.00	0.72 ± 0.07	0.99 ± 0.00	0.74 ± 0.06

Datasets Experiments are conducted on three widely used benchmark datasets: Indian Pines (IP), acquired by the AVIRIS sensor, consists of a 145×145 image with 200 spectral bands after noise removal, a spatial resolution of 20 m, and 16 land-cover classes. Pavia University (PU), collected by the ROSIS sensor, contains a 610×340 image with 103 spectral bands, a spatial resolution of 1.3 m, and 9 urban classes. Salinas (SA), acquired by AVIRIS, has a size of $512 \times 217 \times 204$, with 54,129 labeled pixels distributed over 16 classes and a spatial resolution of 3.7 m.
Evaluated Models Models are selected based on three criteria: state-of-the-art performance, architectural diversity (CNNs vs. Transformers), and publication timeline. We evaluate SSRN [26], Hamida et al.'s 3D-CNN [25], ViT and SpectralFormer [28], and the recent DSFormer [34]. All models use their original configurations.
Evaluation Protocol All experiments are conducted on an NVIDIA RTX A4000 GPU. Quantitative evaluation relies on a 4-fold cross-validation under both sampling strategies. We report the mean and standard deviation of Overall Accuracy (OA), and F1-score.

For the proposed protocol, the training region is located on the right side of the image and the test region on the left for all datasets. During cross-validation, the validation subset alternates between the right, left, top, and bottom regions of the training area. For random sampling, a 30%/70% train/test split is used. For the proposed protocol, the split is 70%/30% for IP and 60%/40% for PU and SA.
Causal Estimation and Attribution Analysis Causal estimation is performed under both protocols using the Average Mutual Information Effect (AMIE) with three regions of interest (ROI): ROI0 (central pixel), ROI1 (3×3 patch), and ROI2 (5×5 patch). Integrated Gradients [35, 36] are employed to compute attribution maps, which are averaged along the spectral dimension to obtain spatial heatmaps. We put in our repository [37] all the code and the information to reproduce the experiments
Results In Table 1, we report classification performance of all selected models using two splitting strategies at the training stage: Random and Split3, across the three considered datasets. The results consistently show that the choice of data splitting strongly affects model performance.

Random sampling systematically overestimates generalization, with accuracy gaps reaching up to 57% (e.g., SpectralFormer on Indian Pines dataset). While the magnitude varies across architectures and datasets, the effect is substantial for all models.

Following the methodology of prior works [38, 2] that study train/test leakage and the relative model accuracy inflation, we compare Split3 to Random splitting, which is the reference baseline. Such a comparison highlights the impact of leakage on performance, while our study also focuses on how leakage affects model behavior.

In Figure 4,5, and 6, we present the results of causal estimation analysis we performed considering three different (by design) models, namely Vit (Visual Transformer), SSRN (Spectral–Spatial Residual Network) and Hamida et al (based on 3D Convolution). On the left-hand side of the figures, we present the AMIE values varying the intervention size, and on the right-hand side, the average feature attribution heatmaps of the patches. Both for Random and Split3 training strategies.

In general, we observe that small neighborhood interventions (T_{ROI_0} and T_{ROI_1}) provide more

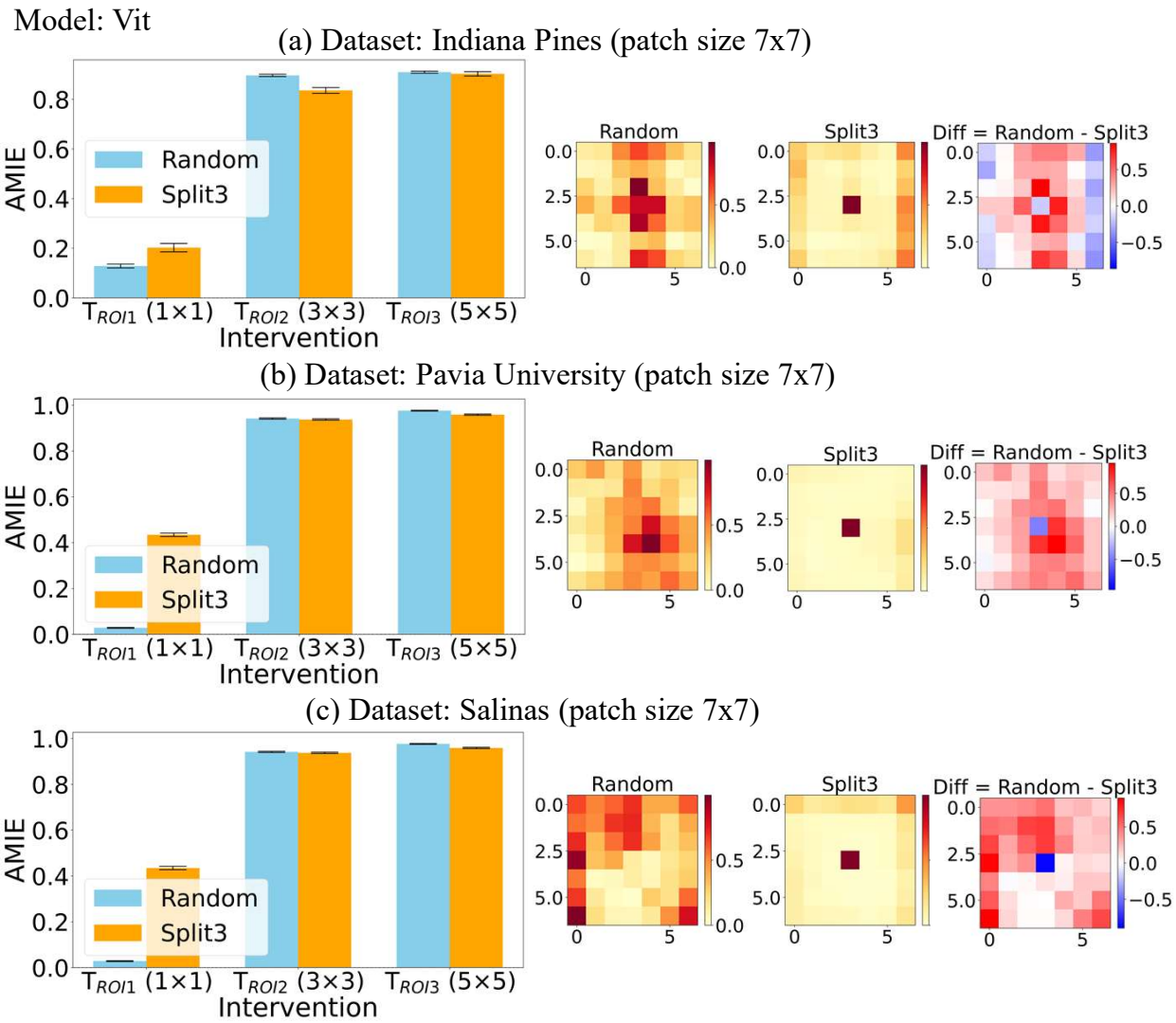

Figure 4: Casual effect estimation with AMIE (left) and feature attribution analysis of HSI classification with the Vit model trained with Random and Split3 strategies. (a) In the Indiana Pines dataset, (b) in the Pavia University Dataset, and (c) in the Salinas dataset.

interpretable results. For larger neighborhoods, AMIE values saturate as the intervention approaches near-complete input ablation, and predictions become dominated by noise, invalidating any conclusions that can be drawn. In this case, when adopting the Split3 sampling, AMIE values are consistently positive across all models, indicating that removing local spatial information decreases the predicted class probability. Such a result provides cues of a stable causal relationship ROI $\rightarrow Y$, although effect magnitudes differ across architectures (weaker for SSRN, stronger for ViT and Hamida 3D-CNN). In contrast, under random sampling, AMIE values at T_{ROI_0} are close to zero, and subsequent interventions rapidly reach saturation, making the causal pathway unidentifiable due to confounding induced by spatial leakage.

Attribution heatmaps qualitatively support these findings: the training with Split3 sampling yields overall spatially coherent attributions concentrated around the patch center as shown in the central heatmaps, whereas random sampling produces more diffuse patterns, consistent with shortcut learning (see left-hand side heatmaps). Such results demonstrate that spatial leakage affects not only reported performance but also the generalizability of learned representations, underscoring the need for leakage-free evaluation protocols in patch-based HSI classification.

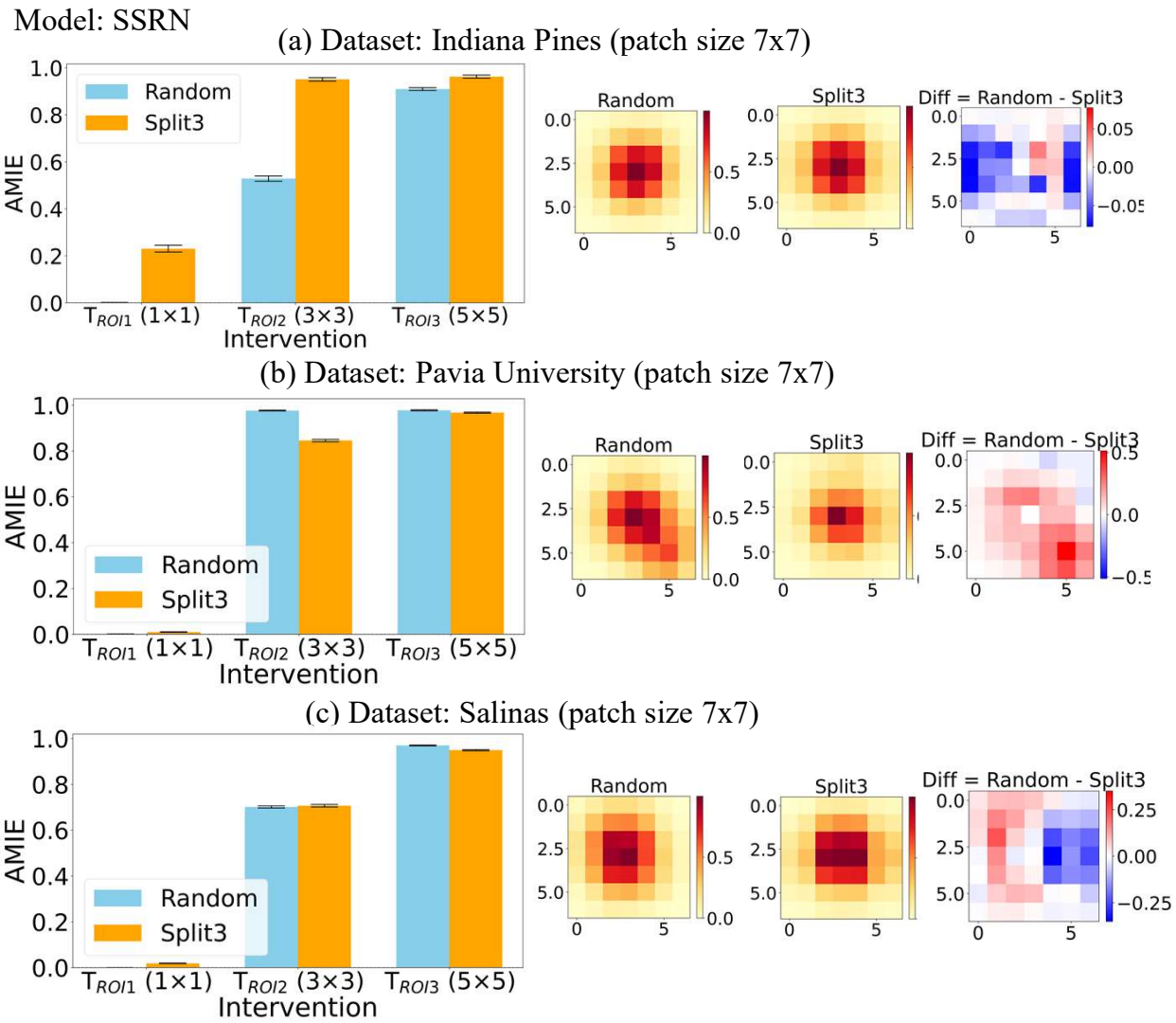

Figure 5: Casual effect estimation with AMIE (left) and feature attribution analysis of HSI classification with the SSRN model trained with Random and Split3 strategies. (a) In the Indiana Pines dataset, (b) in the Pavia University Dataset, and (c) in the Salinas dataset.

5. Conclusions

In this paper, we revisited the benchmark of patch-based hyperspectral image classification from a causal estimation perspective. Our analysis reveals that in the absence of leakage, models rely more consistently on the Region of Interest to support their predictions, whereas under random sampling, the stated causal link becomes obscured by confounding spatial correlations. These findings highlight that reliable assessment of HSI classifiers requires both leakage-free evaluation and causal scrutiny of learned representations. Beyond improving benchmarking practices, this work calls for a rethinking of patch-based HSI classification toward models and protocols that favor causal fidelity over shortcut-driven performance.

Acknowledgments

This work is supported by the "EDEM: Explaining Deep Learning Models of Satellite Time Series for the Agritech domain" project funded by l'Agence Nationale de la Recherche (ANR), EDEM project

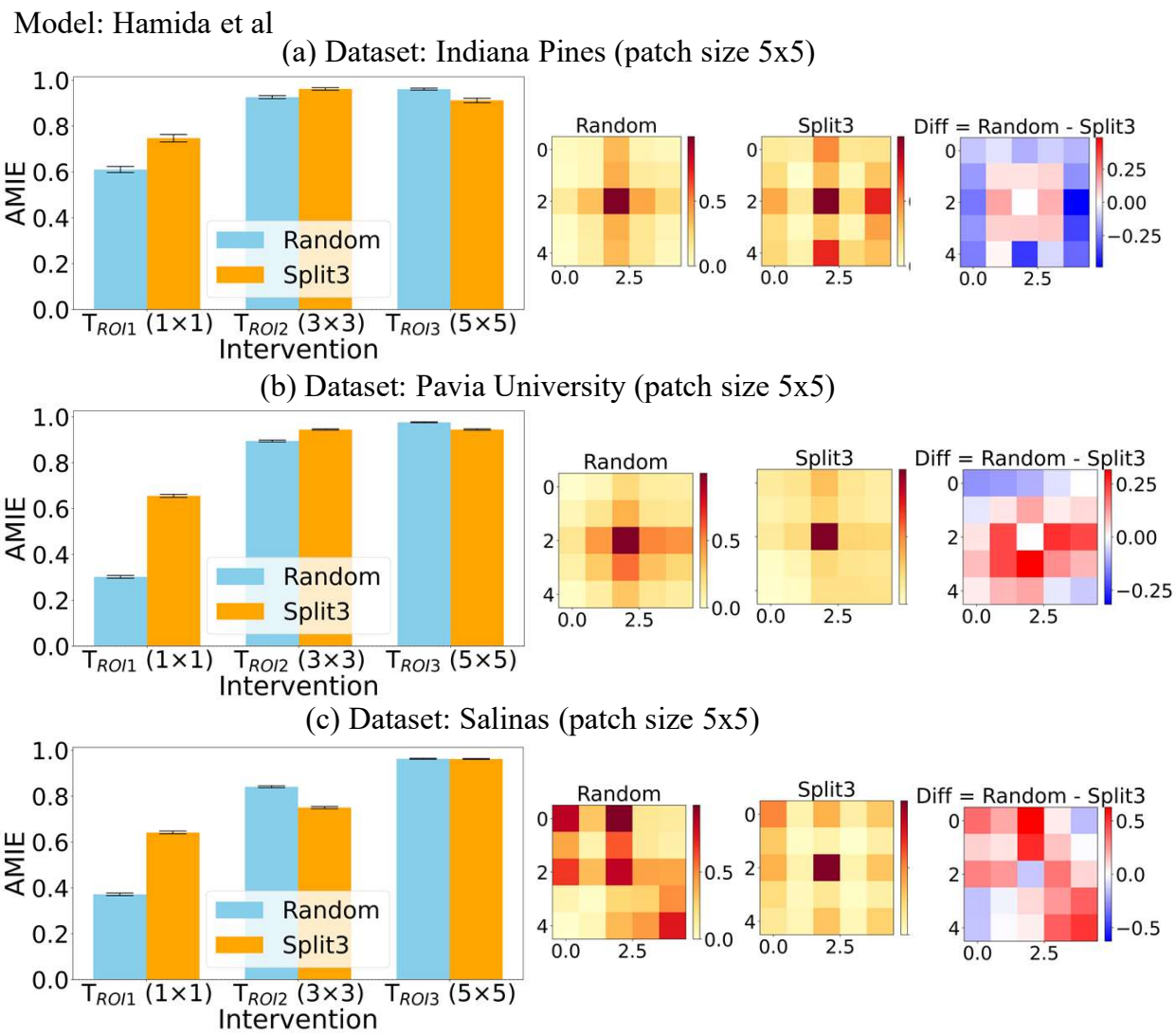

Figure 6: Casual effect estimation with AMIE (left) and feature attribution analysis of HSI classification with the Hamida et al model trained with Random and Split3 strategies. (a) In the Indiana Pines dataset, (b) in the Pavia University Dataset, and (c) in the Salinas dataset.

ANR-24-CE23-6449.

References

[1] M. Ahmad, S. Distifano, A. M. Khan, M. Mazzara, C. Li, H. Li, J. Aryal, Y. Ding, G. Vivone, D. Hong, A Comprehensive Survey for Hyperspectral Image Classification: The Evolution from Conventional to Transformers and Mamba Models, Neurocomputing 644 (2025) 130428. URL: http://arxiv.org/abs/2404.14955. doi:10.1016/j.neucom.2025.130428, arXiv:2404.14955 [cs].

[2] Z. Zhen, L. J. Quackenbush, S. V. Stehman, L. Zhang, Impact of training and validation sample selection on classification accuracy and accuracy assessment when using reference polygons in object-based classification, International Journal of Remote Sensing 34 (2013) 6914–6930. URL: https://doi.org/10.1080/01431161.2013.810822. doi:10.1080/01431161.2013.810822. arXiv:https://doi.org/10.1080/01431161.2013.810822.

[3] J. Zhou, J. Liang, Y. Qian, Y. Gao, L. Tong, On the sampling strategies for evaluation of joint spectral-spatial information based classifiers, in: 2015 7th Workshop on Hyperspectral Image and

Signal Processing: Evolution in Remote Sensing (WHISPERS), IEEE, Tokyo, Japan, 2015, pp. 1–4. URL: http://ieeexplore.ieee.org/document/8075474/. doi:10.1109/WHISPERS.2015.8075474.

[4] J. Liang, J. Zhou, Y. Qian, L. Wen, X. Bai, Y. Gao, On the Sampling Strategy for Evaluation of Spectral-spatial Methods in Hyperspectral Image Classification, IEEE Transactions on Geoscience and Remote Sensing 55 (2017) 862–880. URL: http://arxiv.org/abs/1605.05829. doi:10.1109/TGRS.2016.2616489, arXiv:1605.05829 [cs].

[5] R. Hansch, A. Ley, O. Hellwich, * HSIC - Correct and still wrong: The relationship between sampling strategies and the estimation of the generalization error, IEEE, Fort Worth, TX, 2017, pp. 3672–3675. URL: http://ieeexplore.ieee.org/document/8127795/. doi:10.1109/IGARSS.2017.8127795.

[6] H. Feng, Y. Wang, Z. Li, N. Zhang, Y. Zhang, Y. Gao, Information Leakage in Deep Learning-Based Hyperspectral Image Classification: A Survey, Remote Sensing 15 (2023) 3793. URL: https://www.mdpi.com/2072-4292/15/15/3793. doi:10.3390/rs15153793, publisher: Multidisciplinary Digital Publishing Institute.

[7] K. T. Decker, B. J. Borghetti, A survey of sampling methods for hyperspectral remote sensing: Addressing bias induced by random sampling, Remote Sensing 2025, Vol. 17, Page 1373 17 (2025) 1373. URL: https://www.mdpi.com/2072-4292/17/8/1373/htmhttps://www.mdpi.com/2072-4292/17/8/1373. doi:10.3390/RS17081373.

[8] L. Qu, X. Zhu, J. Zheng, L. Zou, Triple-Attention-Based Parallel Network for Hyperspectral Image Classification, Remote Sensing 13 (2021) 324. URL: https://www.mdpi.com/2072-4292/13/2/324. doi:10.3390/rs13020324.

[9] J. Nalepa, M. Myller, M. Kawulok, Validating Hyperspectral Image Segmentation, IEEE Geoscience and Remote Sensing Letters 16 (2019) 1264–1268. URL: http://arxiv.org/abs/1811.03707. doi:10.1109/LGRS.2019.2895697, arXiv:1811.03707 [cs].

[10] J. Lange, G. Cavallaro, M. Götz, E. Erlingsson, M. Riedel, The Influence of Sampling Methods on Pixel-Wise Hyperspectral Image Classification with 3D Convolutional Neural Networks, 2018, pp. 2087–2090. doi:10.1109/IGARSS.2018.8518671.

[11] N. Karasiak, J.-F. Dejoux, C. Monteil, D. Sheeren, Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing, Machine Learning 111 (2022) 2715–2740. URL: https://doi.org/10.1007/s10994-021-05972-1. doi:10.1007/s10994-021-05972-1.

[12] L. M. Zintgraf, T. S. Cohen, T. Adel, M. Welling, Visualizing deep neural network decisions: Prediction difference analysis, 2017. URL: https://arxiv.org/abs/1702.04595. arXiv:1702.04595.

[13] J. Li, W. Monroe, D. Jurafsky, Understanding neural networks through representation erasure, 2017. URL: https://arxiv.org/abs/1612.08220. arXiv:1612.08220.

[14] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, O. Reblitz-Richardson, Captum: A unified and generic model interpretability library for pytorch, 2020. arXiv:2009.07896.

[15] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019. URL: https://arxiv.org/abs/1910.10045. arXiv:1910.10045.

[16] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, 2016. URL: https://arxiv.org/abs/1602.04938. arXiv:1602.04938.

[17] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, Nature Communications 10 (2019). URL: http://dx.doi.org/10.1038/s41467-019-08987-4. doi:10.1038/s41467-019-08987-4.

[18] M. Graña, M.-A. Veganzones, B. Ayerdi, Pavia university hyperspectral dataset, https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes, 2002. Acquired by the ROSIS sensor over Pavia, Italy. 103 spectral bands; 9 land-cover classes.

[19] G. Hughes, On the mean accuracy of statistical pattern recognizers, IEEE Transactions on Information Theory 14 (1968) 55–63. doi:10.1109/TIT.1968.1054102.

[20] G. Foody, A. Mathur, A relative evaluation of multiclass image classification by support vector

machines, IEEE Transactions on Geoscience and Remote Sensing 42 (2004) 1335–1343. doi:10.1109/TGRS.2004.827257.

[21] W. Hu, Y. Huang, L. Wei, F. Zhang, H. Li, Deep Convolutional Neural Networks for Hyperspectral Image Classification, Journal of Sensors 2015 (2015) 258619. URL: https://onlinelibrary.wiley.com/doi/abs/10.1155/2015/258619. doi:10.1155/2015/258619, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2015/258619.

[22] L. Mou, P. Ghamisi, X. X. Zhu, Deep recurrent neural networks for hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing 55 (2017) 3639–3655. doi:10.1109/TGRS.2016.2636241.

[23] Y. Tarabalka, J. A. Benediktsson, J. Chanussot, Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques, IEEE Transactions on Geoscience and Remote Sensing 47 (2009) 2973–2987. doi:10.1109/TGRS.2009.2016214.

[24] Y. Chen, H. Jiang, C. Li, X. Jia, P. Ghamisi, *** HSIC - Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks, IEEE Trans. Geosci. Remote Sensing 54 (2016) 6232–6251. URL: http://ieeexplore.ieee.org/document/7514991/. doi:10.1109/TGRS.2016.2584107.

[25] A. Ben Hamida, A. Benoit, P. Lambert, C. Ben Amar, 3-d deep learning approach for remote sensing image classification, IEEE Transactions on Geoscience and Remote Sensing 56 (2018) 4420–4434. doi:10.1109/TGRS.2018.2818945.

[26] Z. Zhong, J. Li, Z. Luo, M. Chapman, Spectral–spatial residual network for hyperspectral image classification: A 3-d deep learning framework, IEEE Transactions on Geoscience and Remote Sensing 56 (2018) 847–858. doi:10.1109/TGRS.2017.2755542.

[27] S. K. Roy, G. Krishna, S. R. Dubey, B. B. Chaudhuri, Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification, IEEE Geoscience and Remote Sensing Letters 17 (2020) 277–281. doi:10.1109/LGRS.2019.2918719.

[28] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, J. Chanussot, Spectralformer: Rethinking hyperspectral image classification with transformers, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–15. URL: http://dx.doi.org/10.1109/TGRS.2021.3130716. doi:10.1109/tgrs.2021.3130716.

[29] Y. Xu, D. Wang, L. Zhang, L. Zhang, Dual selective fusion transformer network for hyperspectral image classification, 2025. URL: https://arxiv.org/abs/2410.03171. arXiv:2410.03171.

[30] A. Vandenhoeke, L. Antson, G. Ballesteros, J. Crabbé, M. Shimoni, Explaining the absorption features of deep learning hyperspectral classification models, in: IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium, 2023, pp. 958–961. doi:10.1109/IGARSS52108.2023.10282988.

[31] G. De Lucia, M. Lapegna, D. Romano, Towards explainable ai for hyperspectral image classification in edge computing environments, Computers and Electrical Engineering 103 (2022) 108381. URL: https://www.sciencedirect.com/science/article/pii/S0045790622005985. doi:https://doi.org/10.1016/j.compeleceng.2022.108381.

[32] D. Cheng, Z. Xu, J. Li, L. Liu, K. Yu, T. D. Le, J. Liu, Linking model intervention to causal interpretation in model explanation, Pattern Recognition 173 (2026) 112814. doi:https://doi.org/10.1016/j.patcog.2025.112814.

[33] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, 2017. URL: https://arxiv.org/abs/1703.01365. arXiv:1703.01365.

[34] Y. Xu, D. Wang, L. Zhang, L. Zhang, Dual selective fusion transformer network for hyperspectral image classification, Neural Networks 187 (2025) 107311.

[35] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning, PMLR, 2017.

[36] M. Linardi, S. Dabo, C. Paris, Optimizing deep learning for satellite hyperspectral data: an xai-driven approach to hyperparameter selection, in: IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium, 2025, pp. 7602–7606. doi:10.1109/IGARSS55030.2025.11243437.

[37] Anonymous paper repository, https://anonymous.4open.science/r/all_test-3C47/README.md, 2026.

[38] I. E. Tampu, A. Eklund, N. Haj-Hosseini, Inflation of test accuracy due to data leakage in deep learning-based classification of oct images, Scientific Data 9 (2022) 580.