Efficient Unsupervised Band Selection for Hyperspectral Imagery with Mamba-based Classifier – An In-Depth Comparative Analysis

Anonymous authors
Paper under double-blind review

Abstract

Band selection is a critical step in processing hyperspectral imagery (HSI), reducing input dimensionality to mitigate redundancy, enhance computational efficiency and improve learning accuracy. Efficient unsupervised deep-learning-based band selection methods have recently garnered significant attention due to their strong feature representation capabilities. In existing literature, we observe that there is a broader and more general line of research regarding *feature selection*, which some recent deep learning-based HSI band selection methods have drawn inspiration from. This work concentrates on efficient unsupervised deep-learning-based band selection methods from the standpoint of unifying two research lines: the more general *feature selection* and the more specific *HSI band selection*. Specifically, we conduct an in-depth comparative analysis in terms of downstream classification performance and computation cost, on six state-of-the-art efficient unsupervised HSI band selection methods, of which one does not involve deep learning and the other five do. Classification experiments are carried out using three publicly available remote sensing benchmark datasets, where we incorporate a recent Mamba-based classifier that outperforms the typical SVM substantially in classification accuracy by a ∼10-20% margin. To our best knowledge, this is the first work that puts together and compares the aforementioned efficient unsupervised methods in the context of HSI band selection and employs a Mamba-based classifier in the analysis.

1 Introduction

Remotely-sensed hyperspectral imagery (HSI) presents significant opportunities in Earth observation. HSI captures hundreds of contiguous spectral bands, providing more detailed insights about the imaging scene compared to conventional RGB images, and has become an effective tool in various applications, such as precision agriculture (Ram et al., 2024), mineral detection (Siebels et al., 2020), landscape classification (A & S, 2023), and even medical diagnosis (Wang et al., 2021b). However, the high dimensionality of HSI imposes significant computational burden on data processing and analysis. Therefore, it is crucial to address the reduction of dimensionality in HSI.

Dimensionality reduction in general has several benefits, including reducing experimental costs (Min et al., 2014), enhancing interpretability (Ribeiro et al., 2016), speeding up computation, reducing memory storage, and even improving the generalization of downstream tasks (Chandrashekar & Sahin, 2014). One popular technique for dimensionality reduction is *feature selection*, which involves the identification and retention of the most informative features, as opposed to feature extraction techniques that alter features' semantics by creating new ones in a lower dimensional space. By retaining the original features, researchers can directly relate model outputs to input data, facilitating insights and hypothesis generation. Moreover, in applications where sensing hardware costs or energy consumption are major concerns, such as in IoT devices or sensor-based systems, feature selection can inform the design of simpler and more affordable hardware.

Feature selection algorithms typically assume most features are uninformative and uncorrelated, and the task is to "identify a small, highly discriminative subset" (Kuncheva et al., 2020), e.g., genes associated with

drug response from the entire genome. In the context of HSI, each spectral band is often considered as a feature. On the contrary, most spectral bands individually offer similar amounts of information since they view the same scene but with often-subtle differences in contrast (Blumberg et al., 2022). In other words, most spectral bands correlate strongly and contain significant redundancy. Therefore, existing state-of-the-art feature selection methods that work well in common data modalities such as grayscale or RGB imagery, text, speech etc., might not be as effective when employed on HSI data.

With the advent of miniaturized hyperspectral cameras that are able to be mounted on autonomous drones (Tuohy et al., 2023), the development of more efficient yet more powerful deep learning architectures, and the advancement of GPU-equipped edge hardware, there is a lot of potential in incorporating light-weight deep learning models into the real-time HSI processing pipeline (Dastranj et al., 2025). In this work, we focus on the software algorithmic component of band selection and pixel-wise classification. Specifically, our experiments are designed to address the following two research questions (RQs):

1. How robust are unsupervised, light-weight, autoencoder-based HSI band selection methods on a variety of datasets and classifiers?
2. How can Mamba-based HSI classifiers further benefit from band selection in terms of accuracy and computation cost?

Fig. 1 displays our experiment workflow. We concentrate on unsupervised, light-weight, autoencoder-based architectures for HSI band selection. We select five recent ones from the literature (two of which have yet to be applied for HSI band selection). An efficient state-of-the-art non-deep-learning method is also included in our comparative analysis. We unify and organize their code in Python and PyTorch for a more fair and convenient comparison, analyze their architecture, selection mechanism, classification performance, and computation cost with three distinct hyperspectral remote sensing datasets.

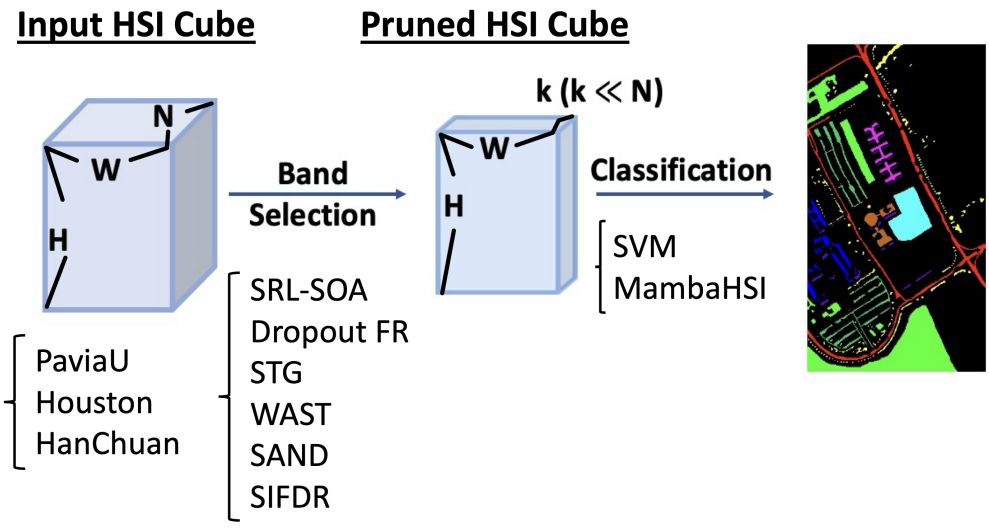


Figure 1: Experiment workflow consisting of two phases – band selection followed by classification. Experiments are carried out with three datasets (see Section 3), six band selection methods (see Section 4.1) and two classifiers (see Section 2.4).

Despite the rapid advancement of deep learning architectures, recent works (Sun et al., 2022; Ahishali et al., 2022; Zhang et al., 2024; Xu et al., 2025) still virtually exclusively employ traditional machine learning classifiers such as support vector machine (SVM) (Melgani & Bruzzone, 2004), random forest (RF) (Ham et al., 2005), and K-nearest neighbors (KNNs) (Jia & Richards, 2005), for evaluating the efficacy of HSI band selection methods. We address this gap by incorporating a recent Mamba-based classifier named MambaHSI (Li et al., 2024) into our comparative analysis and demonstrating their distinct behavior compared to SVM in terms of classification accuracy with respect to the selected band subsets.

Our key contributions are summarized as follows and in Fig. 2:

1. To our best knowledge, this is the first work that puts together and compares efficient unsupervised deep-learning-based methods for HSI band selection from the standpoint of unifying two research lines in literature – the more general *feature selection* and the more specific *HSI band selection*.

2. We are also the first to include a Mamba-based classifier in HSI band selection research and reveal their distinct behaviors compared to the conventional SVM.

3. Inspired by two prior works (Sun et al., 2022; Xu et al., 2025) that design an unsupervised HSI band selection method by incorporating an existing feature selection mechanism into an autoencoder, we follow suit for another two recent feature selection methods, resulting in their first applications in literature for HSI band selection.

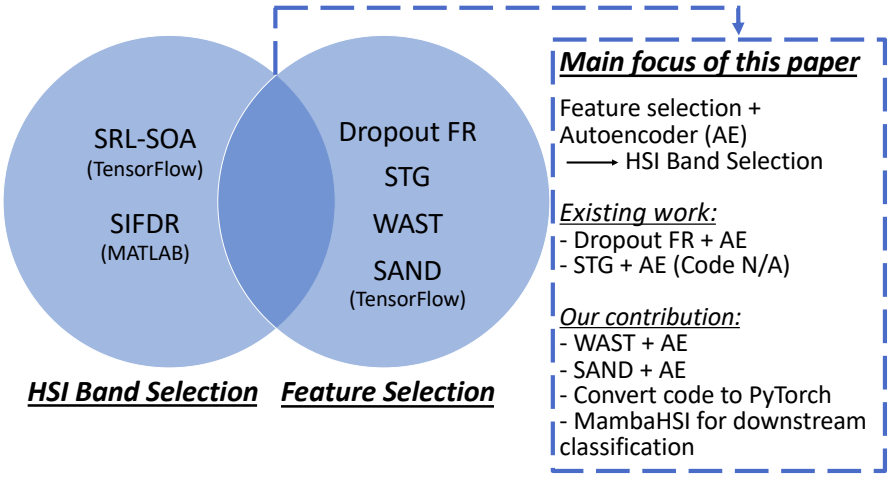


Figure 2: Contributions of this work. We concentrate on autoencoder-based feature selection techniques that can be adapted for HSI band selection. We convert code to Python/PyTorch for a more fair and convenient comparison and include a Mamba-based model in our downstream classification analysis. Text in parentheses indicates either code not available or code implemented in languages/frameworks other than Python/PyTorch.

The remainder of this article is structured as follows: Section 2 presents related work in feature selection, unsupervised HSI band selection and HSI classification. The datasets are detailed in Section 3. Section 4 elucidates six state-of-the-art efficient unsupervised techniques that we study in this work. Section 5 exhibits experimental results and analyses. Finally, Section 6 draws conclusions.

2 Background

2.1 Feature Selection

Feature selection operates either at the instance level (Covert et al., 2023), e.g., identifying different salient parts of different images, or at the population level by selecting across all the instances. For hyperspectral imagery (HSI), spectral band selection is population-wide.

According to literature, feature selection methods are generally divided into three classes: **filter methods**, **wrapper methods**, and **embedded methods**. **Filter methods** attempt to remove irrelevant features prior to learning a predictive model. They rely on a per-feature relevance score (e.g., Laplacian score (He et al., 2005)) based on statistical or information-theoretic measures. While these methods are fast and can handle high-dimensional data, they overlook intricate relationships between features. **Wrapper methods** exploit the performance of a predictive model to evaluate the quality of a subset of features. They require recomputing the model for each subset of features, and thus can be prohibitively computationally expensive although being more effective than filter methods. Examples include SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) , greedy sequential search (Das & Kempe, 2011) and other search-based algorithms (Morales et al., 2021). **Embedded methods** rank features based on metrics intrinsically learned during model training, seamlessly integrating feature selection into the learning process. Examples include feature importance for tree-based algorithms (Breiman, 2001), Recursive Feature Elimination for Support Vector Machine (RFE-SVM) (Guyon et al., 2002), and deep learning techniques (Simonyan et al., 2013; Wang et al., 2014). Such methods enable an automatic selection of relevant features during training and can effectively handle non-trivial relationships in data.

2.2 Unsupervised HSI Band Selection

A multitude of unsupervised HSI band selection approaches have been proposed to tackle the band redundancy problem, and according to literature, they are often categorized into four groups: **ranking-based**, **clustering-based**, **searching-based**, and **deep learning-based**. **Ranking-based methods** (Chang et al., 1999; Jia et al., 2016) rank the significance of each band based on some statistical characteristics, e.g., structural similarity (SSIM) (Xu et al., 2021). However, the correlation between selected bands are very high, implying considerable redundancy. **Clustering-based methods** (Sun et al., 2015; Wang et al., 2018; 2021a) aim to reduce the correlation between selected bands compared to ranking-based methods. In general, clustering-based methods group relevant bands based on some similarity measures and then select the ones closest to the center from each cluster; thus significantly reducing redundancy among the chosen bands. However, this strategy may ignore some representative bands if they are grouped into the same cluster. Besides, the clustering process is very sensitive to noise. **Searching-based methods** (Morales et al., 2021; Wang et al., 2020) use specific search strategies and objective functions, such as maximum information and minimum redundancy (MIMR) (Feng et al., 2016) and maximum information and minimum noise (MIMN) (Chen et al., 2020), to directly find an optimal band subset through iterative exploration and evaluation. However, their computation burden is rather high due to numerous iterations. The majority of **deep learning-based** methods utilize the autoencoder (AE) architecture (Feng et al., 2021; Cai et al., 2020; Ahishali et al., 2022; Liu et al., 2022), and demonstrate superiority over their non-deep-learning counterparts, for their capabilities to learn non-linear dependencies among input features (Li et al., 2017).

2.3 Unifying Feature Selection and HSI Band Selection

As displayed in Fig. 3, the categorization of *Feature Selection* and *HSI Band Selection* can be unified. Quite straightforwardly, ranking and searching-based methods for HSI band selection fall under filter and wrapper methods, respectively. Clustering and deep-learning-based methods, on the other hand, belong to the embedded feature selection category as the features are ranked upon the clustering/learning process. Given the pervasive adoption of deep learning in recent years, this work concentrates on embedded feature selection techniques with regard to deep learning. Specifically, we are interested in autoencoders that efficiently identify a subset of the most informative features and simultaneously learn a neural network to reconstruct the input data from the selected features.

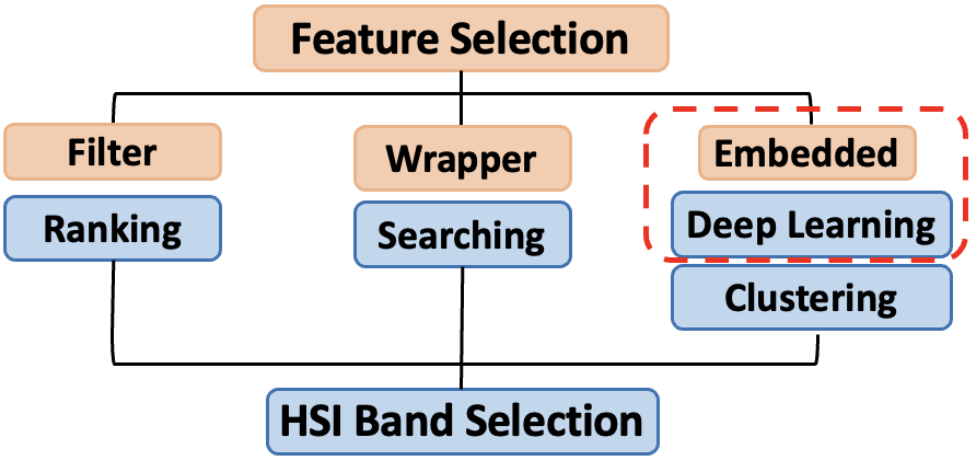


Figure 3: The unification of categorization of two lines of research – the more general *feature selection* and the more specific *HSI band selection*. In this work, we concentrate on efficient embedded techniques using autoencoders for HSI band selection.

2.4 HSI Classifiers

Traditional machine learning classifiers, such as support vector machine (SVM) (Melgani & Bruzzone, 2004), random forest (RF) (Ham et al., 2005), and K-nearest neighbors (KNNs) (Jia & Richards, 2005), treat each spatial pixel as an independent sample and thus ignore the spatial context, yielding suboptimal results. Likewise, there are inherent limitations in mainstream deep-learning based HSI classifiers, which are CNN or Transformer-based. CNN-based models are constrained by their local receptive fields, hindering their ability to model long-range dependencies. Although Transformer-based models show superior performance for global modeling, their self-attention mechanism scales quadratically with sequence length. In contrast, Mamba (Gu & Dao, 2024), an emerging architecture based on selective state space models (SSMs), is adept at continuous long-sequence data analysis while maintaining linear complexity in terms of sequence length (Gu et al., 2021). Moreover, due to their limitations, CNN and Transformer-based models process HSI data in patches, whereas Mamba-based models are capable of taking the entire image as input and process at a fine-grained pixel level.

MambaHSI (Li et al., 2024) is a recent Mamba-based HSI classifier that we employ in our experiments. It's backbone contains two main components: Spatial Mamba Block (SpaMB) and Spectral Mamba Block (SpeMB), to extract discriminative spatial and spectral features. Table 1 summarizes the function of the two Mamba blocks from the perspective of processing a 1-D sequence, where the HSI data cube has a spatial size of $H \times W$, and C spectral bands. For SpeMB, the full spectrum is divided into G equal groups ($G = 4$ is used in the experiments). Batch size 1 indicates the entire HSI image is processed at once without being broken into patches.

SVM is the default classifier in most HSI band selection research (Sun et al., 2022; Ahishali et al., 2022; Zhang et al., 2024; Xu et al., 2025); thus we include it in our experiments and analysis.

Table 1: Summary of the function of the two Mamba blocks in MambaHSI from the perspective of 1-D sequence processing

	Batch Size	**Sequence Length**	**Embedding Dimension**
Spatial Mamba Block	1	$H \times W$	C
Spectral Mamba Block	$1 \times H \times W$	G	$C \div G$

3 Datasets – Land Cover Land Use (LCLU) Classification

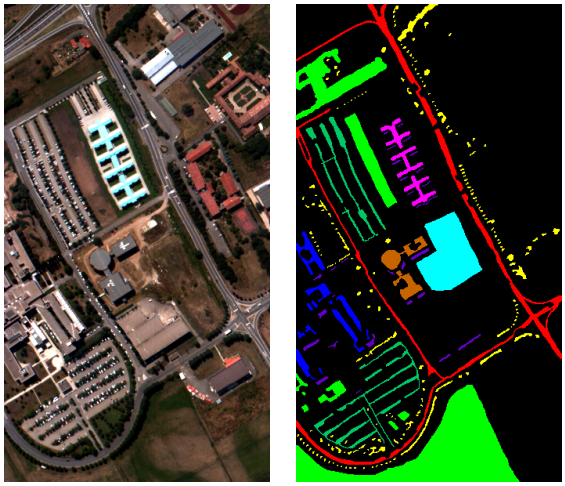
Table 2: Summary of the three studied hyperspectral remote sensing datasets

Dataset	Spectral Bands	Spectral Range	Total Pixels	Classes	Spatial Resolution
PaviaU	103	430-860 nm	610 × 340	9	1.3 m
Houston	144	380-1050 nm	349 × 1905	15	2.5 m
HanChuan	274	400-1000 nm	1217 × 303	16	0.109 m

Following Li et al. (2024), we select three widely used and diverse hyperspectral benchmark datasets, which include both urban and agricultural scenes and have different spatial resolutions. Table 2 summarizes the details of each dataset. For every dataset, during band selection (phase 1), 1000 pixels are randomly selected from the entire HSI regardless of whether the selected pixel has a class label or not; during classification (phase 2), for each class, 30 pixels are randomly selected as training samples, 10 pixels are randomly selected as validation samples, and the remaining labeled pixels are used for testing.

3.1 Pavia University (PaviaU)

The image (Fig. 4) was acquired over University of Pavia in 2002 with the Reflective Optics System Imaging Spectrometer (ROSIS) sensor, consisting of 103 spectral bands covering the 430-860nm spectral range, with a spatial size of 610 × 340 pixels and a 1.3m ground sampling distance (GSD). The image contains 9 classes.



(a) (b)

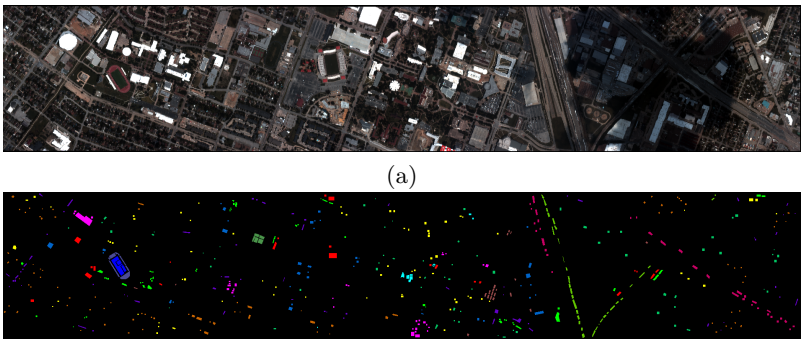
Figure 4: (a) False color image and (b) ground truth map of Pavia University.

3.2 Houston (Debes et al., 2014)

The image (Fig. 5) was acquired over the University of Houston campus and its neighboring regions with the ITRES CASI 1500 HS imager, consisting of 144 spectral bands covering the 380-1050nm spectral range, with a spatial size of 349 × 1905 pixels and a 2.5m ground sampling distance (GSD) . The image contains 15 classes. It was provided by the 2013 IEEE Geoscience and Remote Sensing Society (GRSS) data fusion contest.

3.3 WHU-Hi-HanChuan (HanChuan) (Zhong et al., 2020)

The image (Fig. 6) was acquired from 17:57 to 18:46 on June 17, 2016, in Hanchuan, Hubei province, China, with an 17-mm focal length Headwall Nano-Hyperspec imaging sensor mounted on a Leica Aibot

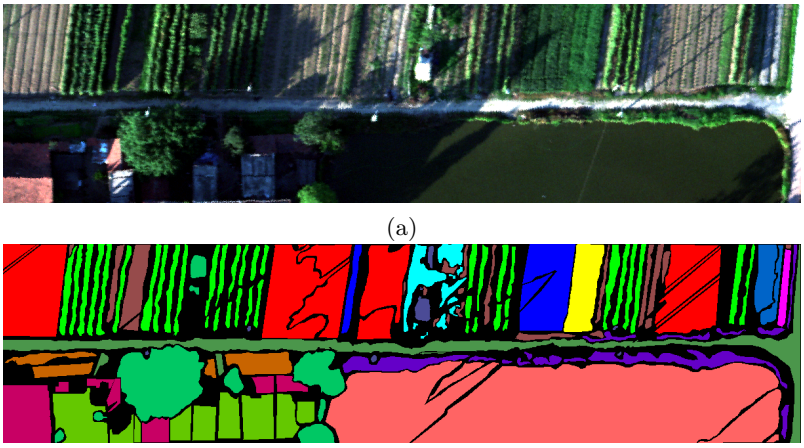


(a)

(b)

Figure 5: (a) False color image and (b) ground truth map of Houston.

X6 UAV V1 platform flying at an altitude of 250m. During data collection, the weather was clear and cloudless, the temperature was about 30°C, and the realtive humidity was about 70%. The studied area is a rural–urban fringe zone with buildings, water and cultivated land including seven crop species: strawberry, cowpea, soybean, sorghum, water spinach, watermelon, and greens. The image consists of 274 bands covering the 400-1000nm spectral range, and has a spatial size of 1217 × 303 pixels and a 0.109m ground sampling distance (GSD). The image contains 16 classes. Notably, since the first two bands are entirely zeros, only the remaining 272 bands are used for the experiments. Moreover, since this dataset was acquired during the afternoon when the solar elevation angle was low, there are many shadow-covered areas in the image.



(a)

(b)

Figure 6: (a) False color image and (b) ground truth map of WHU-Hi-HanChuan. (Rotated 90 degrees counterclockwise)

4 Experiments

The classification experiments were executed on a workstation equipped with NVIDIA RTX 6000 GPUs and AMD EPYC 7282 16-Core Processors. The algorithms are implemented with PyTorch 2.7.0 and Python 3.9.21.

4.1 Compared Methods

This section briefly describes the mechanism of six state-of-the-art efficient unsupervised techniques that we study in this work, including five deep-learning light-weight autoencoder-based methods (one specifically for

HSI band selection and the other four originally for the more general *feature selection*), and one non-deep-learning method. We denote N as the total number of spectral bands and k as the number of selected bands, where $k \ll N$.

4.1.1 Self-Representation Learning with Sparse 1D-Operational Autoencoder (SRL-SOA) (Ahishali et al., 2022)

This method assumes that each HSI band can be represented by a linear combination of all other bands. The encoder part consists of Q 1D convolutional layers and outputs a representation matrix $\mathbf{A}$ of size $N \times N$. **Band importance is determined by** the row sum of $\mathbf{A}$ – the larger the sum, the more important the band. Diagonal entries of $\mathbf{A}$ are zeroed out to prevent trivial solutions where each band is represented by itself. ℓ_1 regularization is applied on $\mathbf{A}$ to impose sparsity. Mathematically, the encoder can be described as the following function:

$$\mathbf{y}^{(p)} = \sigma(\sum_{q=1}^{Q}(\mathbf{x})^q * \mathbf{w}_q^{(p)} + b_q^{(p)}) \tag{1}$$

where σ is the activation function (hyperbolic tangent) and p indicates the p^{th} filter. We adopt $Q = \{1, 3\}$ in our experiments, where a larger Q indicates a higher degree of nonlinearity.

4.1.2 Dropout Feature Ranking (FR) (Chang et al., 2017)

This method employs the mechanism of the widely used overfitting prevention technique: Dropout (Srivastava et al., 2014), where a multiplicative Bernoulli noise is injected into each hidden neuron within a neural network. However, instead of having preset, fixed dropout rates, they adopt Variational Dropout (Maeda, 2014), where the dropout rates are learned and optimized. This Variational Dropout regularization is applied to the input layer to perform feature ranking. **Band importance is determined by** the optimized dropout rates – the lower the rate, the more important of a band. The Bernoulli discrete variables are optimized through Concrete relaxation (Jang et al., 2016; Maddison et al., 2016) – during training, the temperature of the concrete selector layer is gradually decreased, which encourages a user-specified number of discrete features to be learned. Xu et al. (2025) have applied Dropout FR to unsupervised HSI band selection.

4.1.3 Feature Selection using Stochastic Gates (STG) (Yamada et al., 2020)

This method was primarily proposed to serve as a feature selection method for non-linear functions, just as the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) is to linear functions. Specifically, they approximate the ℓ_0 norm of features (i.e., the count of selected features) via Gaussian-based relaxation of the Bernoulli distribution, so that gradient descent based optimization can be used. Gaussian-based relaxation is preferred over the heavy-tailed logistic distribution used in the Concrete relaxation (Jang et al., 2016; Maddison et al., 2016) for better consistency in the selected set of features. Mathematically, the Stochastic Gate (STG, i.e., the relaxed Bernoulli variable) is defined as $z_d = \max(0, \min(1, \mu_d + \epsilon_d))$, where ϵ_d is drawn from $\mathcal{N}(0, \sigma^2)$, σ is a fixed, preset hyperparameter, μ_d is learned and $d \in [N]$. This approximation can be viewed as a clipped, mean-shifted, Gaussian random variable. In principle, $\hat{z_d} = \max(0, \min(1, \mu_d))$ **determines band importance**, where the values of $\hat{z_d}$ converge to either 0 or 1. Sun et al. (2022) have applied STG to unsupervised HSI band selection – their method is named "stochastic gate-based autoencoder (SGAE)", but their code is not publicly available.

4.1.4 Where to Pay Attention in Sparse Training for Feature Selection? (WAST) (Sokar et al., 2022)

This method jointly optimizes an autoencoder's weights and its sparse topology (i.e., connectivity) during training to quickly pay attention to informative features. A preset sparsity level is kept fixed throughout training, and the sparse topology is optimized so that the sparse connections are gradually redistributed to the most informative features (i.e., neurons at the input layer). The authors define neuron importance based on the magnitude of the gradient with respect to reconstruction loss and the connecting weights. **Band**

importance is determined by the neuron importance at the input layer. Thus far this method has not been experimented with HSI band selection in the literature.

4.1.5 SAND: One-Shot Feature Selection with Additive Noise Distortion (Pad et al., 2025)

This method introduces a novel, non-intrusive feature selection layer such that no alteration to the loss function (i.e., no additional regularization term) is required during training. The layer is mathematically elegant and can be fully described by

$$\tilde{x_i} = a_i x_i + (1 - a_i) z_i \tag{2}$$

where x_i is the input feature, z_i is a zero-mean Gaussian noise, and a_i is a trainable parameter such that $\sum_{i=1}^{N} a_i^2 = k$. This formulation induces an automatic clustering effect, driving k of the a_i's to 1 (selecting informative features) and the rest to 0 (discarding redundant ones) via weighted noise distortion and normalization of a_i's. **Band importance is determined by** the magnitude of a_i. Thus far this method hasn't been applied to HSI band selection. Inspired by (Sun et al., 2022) and (Xu et al., 2025), we incorporate the SAND selection layer into an autoencoder structure.

4.1.6 A Real-Time Unsupervised Hyperspectral Band Selection via Spatial-Spectral Information Fusion-Based Downscaled Region (SIFDR) (Zhang et al., 2024)

This method is the only non-deep-learning method included in this work. It first employs bicubic interpolation to downscale the HSI spatially by a factor of $\frac{1}{8} \times \frac{1}{8}$. This has three benefits: 1.) mitigating noise speckles, 2.) computational time and memory usage are substantially reduced, and 3.) the interpolation process involves weighted calculations of neighboring pixels, which implicitly incorporates spatial context in the downscaled image. Then, a vector of length N that assigns an importance weight to each band is generated based on constrained energy. Finally, to generate a more representative band subset with less redundancy, this vector is refined with the aid of the Euclidean distance between bands that are normalized with the hyperbolic tangent function to suppress noise for their saturation property. **Band importance is determined by** the refined weight vector.

4.2 Evaluation Metrics

Following literature conventions (Ahishali et al., 2022; Xu et al., 2025), we use the classification results to evaluate the efficacy of different band selection techniques described in Section 4.1. There are three metrics: overall accuracy (OA), average per-class accuracy (AA), and the Cohen's kappa coefficient (Kappa). We report the mean and standard deviation of results over five runs with different seeds. The larger the mean and the smaller the standard deviation, the better the performance.

5 Results and Discussion

Fig. 7 depicts the overall classification accuracy of applying different band selection algorithms with various band budgets. Comparing the classifiers, it can be clearly observed that MambaHSI outshines SVM significantly, especially with the more challenging dataset WHU-Hi-HanChuan. Moreover, in stark contrast to SVM, MambaHSI can handle randomly generated band subsets just as well as those generated by sophisticated selection algorithms. Also, as opposed to MambaHSI's accuracies improving with the number of bands, SVM's curves are significantly flatter. Notably, when only 5 bands are selected, SVM tends to outperform MambaHSI.

For the Pavia University dataset, which is the easiest out of the three, each band selection method exhibits similar levels of accuracy. Interestingly, with the slightly harder Houston dataset, the band selection algorithms demonstrate consistent behavior across the two classifiers, where two algorithms – WAST and SIFDR, are the notable two low performers when compared against other algorithms that show similar performance. It is the most challenging WHU-Hi-HanChuan dataset that the band selection algorithms show notably

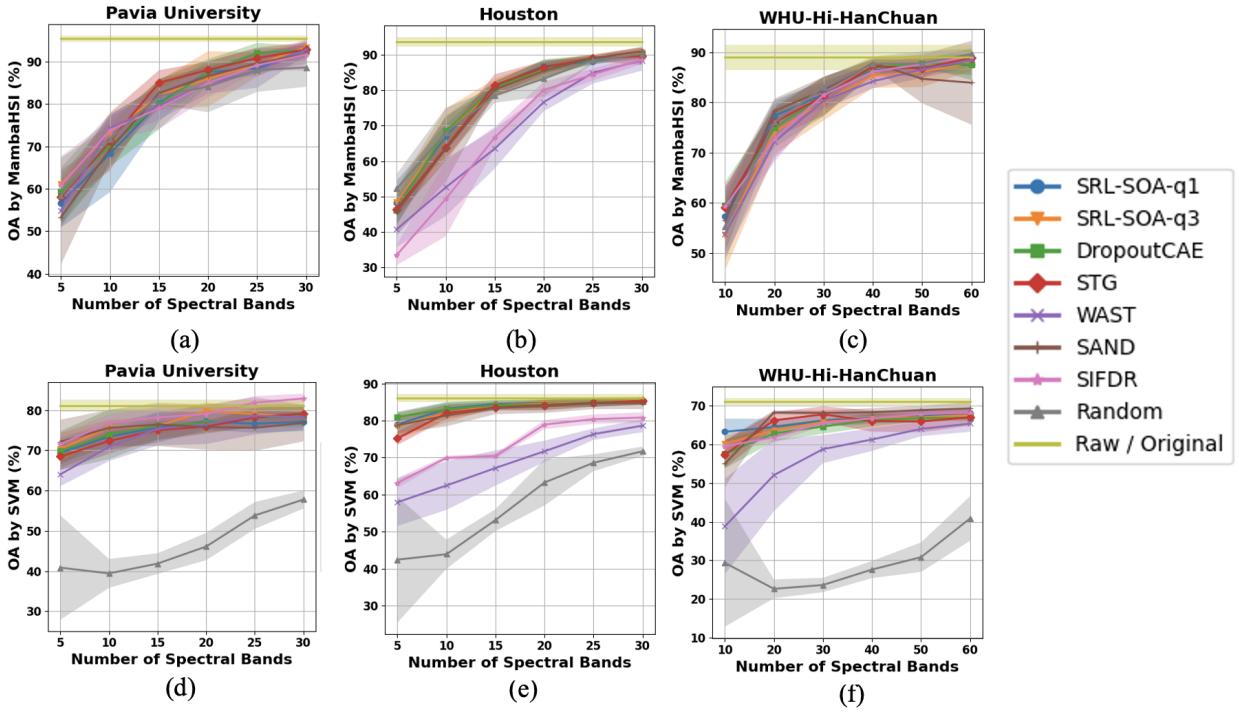


Figure 7: Overall accuracy of three HSI datasets: Pavia University (a)(d), Houston (b)(e), and WHU-Hi-HanChuan (c)(f) for various band selection methods using two classifiers: MambaHSI (containing only the Spatial Mamba Block (SpaMB)) (a)-(c) and SVM (d)-(f). (Random: randomly select a subset of bands; Raw/Original: Full HSI data)

different result patterns between the two classifiers. While the accuracy curves produced by MambaHSI are tangled together, WAST works significantly less effective with SVM.

Table 3: Classification results of different band selection algorithms for Pavia University dataset (30 bands selected out of 103 total bands [∼29% selected])

Method	**MambaHSI**			**SVM**		
	OA(%)	AA(%)	Kappa(%)	OA(%)	AA(%)	Kappa(%)
Random	88.64 ± 4.55	91.76 ± 2.27	85.36 ± 5.74	57.77 ± 2.23	68.89 ± 2.43	48.01 ± 2.31
SRL-SOA (Q=1)	92.77 ± 2.16	$\underline{94.29 \pm 0.72}$	90.56 ± 2.71	77.13 ± 2.34	83.11 ± 1.34	70.83 ± 2.90
SRL-SOA (Q=3)	$\mathbf{93.44 \pm 1.22}$	$\mathbf{94.40 \pm 0.87}$	$\mathbf{91.38 \pm 1.57}$	78.29 ± 1.07	83.58 ± 1.31	72.19 ± 1.41
Dropout FR	92.65 ± 1.07	93.68 ± 0.67	90.35 ± 1.37	78.61 ± 2.75	$\underline{83.87 \pm 1.45}$	72.59 ± 3.38
STG	92.83 ± 1.76	94.04 ± 0.18	90.61 ± 2.24	$\underline{79.13 \pm 1.96}$	83.27 ± 0.94	$\underline{73.06 \pm 2.43}$
WAST	$\underline{92.89 \pm 1.33}$	93.49 ± 0.93	$\underline{90.66 \pm 1.70}$	78.25 ± 3.06	83.67 ± 1.45	72.21 ± 3.63
SAND	91.79 ± 2.32	93.77 ± 0.35	89.30 ± 2.94	76.82 ± 4.61	82.74 ± 1.66	70.46 ± 5.33
SIFDR	91.69 ± 3.74	93.34 ± 1.77	89.20 ± 4.69	$\mathbf{82.88 \pm 1.20}$	$\mathbf{86.13 \pm 0.49}$	$\mathbf{77.80 \pm 1.44}$
All Bands (SpaMB-ONLY)	95.39 ± 0.70	95.81 ± 0.48	93.93 ± 0.89	-	-	-
All Bands (Full Model)	96.07 ± 0.41	96.04 ± 0.33	94.80 ± 0.53	81.08 ± 1.55	85.31 ± 0.92	75.62 ± 1.97

Tables 3, 4 and 5 record the performance of each studied dataset respectively. We record the performance using 30 selected bands for Pavia University (103 bands total) and Houston (144 bands total) datasets, and 60 selected bands for WHU-Hi-HanChuan dataset (272 bands total). The best values are highlighted in bold, and the second-best values are underlined. The first row of each table records the performance from a randomly generated band subset, which we intend to serve as a lower bound. For the upper bound, we also record the performance using the original HSI data containing all the bands.

Table 4: Classification results of different band selection algorithms for Houston dataset (30 bands selected out of 144 total bands [∼21% selected])

Method	MambaHSI			SVM		
	OA(%)	AA(%)	Kappa(%)	OA(%)	AA(%)	Kappa(%)
Random	90.21 ± 1.35	91.62 ± 1.07	89.41 ± 1.46	71.69 ± 1.25	73.34 ± 1.26	69.42 ± 1.35
SRL-SOA (Q=1)	$\underline{90.85 \pm 0.56}$	$\underline{92.19 \pm 0.54}$	$\underline{90.10 \pm 0.60}$	85.17 ± 0.87	85.44 ± 0.69	83.95 ± 0.94
SRL-SOA (Q=3)	90.61 ± 1.40	92.15 ± 1.12	89.84 ± 1.51	$\underline{85.19 \pm 0.97}$	85.42 ± 0.83	$\underline{83.98 \pm 1.04}$
Dropout FR	89.92 ± 1.10	91.34 ± 1.01	89.10 ± 1.18	$\underline{85.19 \pm 0.97}$	$\underline{85.59 \pm 0.94}$	83.98 ± 1.05
STG	89.56 ± 2.45	91.24 ± 1.97	88.71 ± 2.64	$\mathbf{85.32 \pm 0.98}$	$\mathbf{85.66 \pm 0.98}$	$\mathbf{84.12 \pm 1.06}$
WAST	88.36 ± 2.85	90.15 ± 2.41	87.41 ± 3.08	78.63 ± 1.75	79.52 ± 1.84	76.91 ± 1.89
SAND	$\mathbf{90.91 \pm 1.39}$	$\mathbf{92.41 \pm 1.07}$	$\mathbf{90.17 \pm 1.50}$	84.74 ± 1.64	85.26 ± 1.59	83.49 ± 1.78
SIFDR	88.64 ± 1.20	90.47 ± 0.96	87.71 ± 1.30	80.82 ± 1.40	81.91 ± 1.25	79.25 ± 1.51
All Bands (SpaMB-ONLY)	93.60 ± 1.35	94.73 ± 1.16	93.08 ± 1.46	-	-	-
All Bands (Full Model)	94.30 ± 1.00	95.21 ± 0.92	93.84 ± 1.09	86.01 ± 1.13	86.32 ± 1.07	84.87 ± 1.22

Table 5: Classification results of different band selection algorithms for WHU-Hi-HanChuan (60 bands selected out of 272 total bands [∼22% selected])

Method	MambaHSI			SVM		
	OA(%)	AA(%)	Kappa(%)	OA(%)	AA(%)	Kappa(%)
Random	89.65 ± 1.05	89.24 ± 1.59	87.96 ± 1.21	40.83 ± 5.75	37.53 ± 3.81	34.37 ± 5.75
SRL-SOA (Q=1)	87.51 ± 2.04	87.06 ± 1.98	85.49 ± 2.35	68.11 ± 1.30	62.08 ± 0.71	63.41 ± 1.42
SRL-SOA (Q=3)	87.84 ± 1.52	87.08 ± 1.87	85.87 ± 1.75	67.48 ± 0.60	61.32 ± 0.61	62.68 ± 0.68
Dropout FR	87.55 ± 3.07	86.66 ± 3.47	85.53 ± 3.55	66.78 ± 0.75	61.39 ± 1.04	61.95 ± 0.80
STG	$\underline{88.91 \pm 0.48}$	$\underline{88.60 \pm 1.14}$	$\underline{87.11 \pm 0.55}$	66.97 ± 2.13	61.08 ± 1.60	62.15 ± 2.29
WAST	88.43 ± 1.34	88.27 ± 1.22	86.56 ± 1.55	65.34 ± 2.09	59.17 ± 1.98	60.30 ± 2.28
SAND	83.92 ± 8.34	83.76 ± 8.18	81.47 ± 9.38	$\mathbf{69.28 \pm 1.64}$	$\mathbf{63.40 \pm 1.05}$	$\mathbf{64.73 \pm 1.76}$
SIFDR	$\mathbf{89.15 \pm 0.25}$	$\mathbf{88.86 \pm 1.14}$	$\mathbf{87.38 \pm 0.28}$	$\underline{68.19 \pm 1.03}$	$\underline{62.30 \pm 0.42}$	$\underline{63.53 \pm 1.08}$
All Bands (SpaMB-ONLY)	88.98 ± 2.49	88.17 ± 2.91	87.16 ± 2.91	-	-	-
All Bands (Full Model)	89.87 ± 1.88	90.01 ± 2.22	88.22 ± 2.16	71.00 ± 0.90	65.85 ± 0.38	66.69 ± 0.94

Overall, MambaHSI outperforms SVM ∼10-20% in accuracy depending on the dataset. Besides, on every dataset, MamabaHSI paired with the least performant band selector still outshines SVM when all bands are in use. There doesn't exist a single best band selection algorithm that works robustly well across distinct datasets and classifiers. One particular band selection method can be the best on one dataset and a notable low performer on the other. Additionally, as the dataset gets more complicated with more classes and more noise, for MambaHSI, the studied band selection methods are not as effective anymore and even perform worse than randomly selecting bands by chance. Particularly for the WHU-Hi-HanChuan dataset, none of the band selection methods exceed the random selector.

5.1 Architectural Analysis of Deep-Learning-Based Band Selection Methods

Table 6: Summary of architecture information for deep-learning-based band selection algorithms

Method	Encoder Architecture	Encoder Output / Decoder Input	Decoder Architecture	Decoder Output	Activation Function
SRL-SOA	Conv 1D layer(s)	representation matrix	-	$\hat{X}$	Tanh
Dropout FR	Concrete selector layer	selected bands	3 FC layers		ReLU
STG	Stochastic gate layer				
SAND	SAND layer				
WAST	2 FC layers	$\hat{X}$	-	-	sigmoid

Table 6 presents architectural information for autoencoder-based band selection algorithms. $\hat{X} \in R^{H \times W \times N}$ denotes the reconstructed HSI data cube, where N is the total number of spectral bands.

The encoder of SRL-SOA (Ahishali et al., 2022) consists of one or more 1D convolutional layers with a kernel size of 3 and both `in_channels` and `out_channels` equal N. The hyperbolic tangent activation function is applied once after all the convolutional layers and before the matrix is constructed. The encoder of WAST (Sokar et al., 2022) consists of two fully connected (FC) layers with `(in_features, out_features)` equal `(N, 200)` and `(200, N)` respectively. The sigmoid activation function is applied after the first FC layer.

Dropout feature ranking (FR) (Chang et al., 2017), stochastic gates (STG) (Yamada et al., 2020), and one-shot feature selection with additive noise distortion (SAND) (Pad et al., 2025) are originally feature selection mechanisms that can be plugged into any model architecture. (Xu et al., 2025) and (Sun et al., 2022) have employed Dropout FR and STG, respectively, as the encoder layer of an autoencoder structure for HSI band selection. Inspired by them, we follow suit for SAND as well. The decoder consists of three FC layers with `(in_features, out_features)` equal `(k, 128)`, `(128, 256)` and `(256, N)` respectively, where `k` is the number of selected bands. ReLU is applied after the first two FC layers.

5.2 Efficiency Analysis

5.2.1 Phase 1: Band Selection

Table 7: Comparison of computation cost for each band selection algorithm on the WHU-Hi-HanChuan dataset for selecting 60 bands

Method	FLOPs	Params	Training Time (s)	Test Time (s)
SRL-SOA (Q=1)	8.88M	222.22K	64.63 ± 1.21	0.066
SRL-SOA (Q=3)	22.22M	666.67K	71.03 ± 0.19	0.080
Dropout FR	3.31M	111.01K	17.99 ± 0.15	-
STG	3.31M	111.01K	21.90 ± 0.54	-
WAST	3.26M	109.27K	15.48 ± 0.66	-
SAND	3.31M	111.01K	16.08 ± 0.13	-
SIFDR	-	-	1.74 ± 0.17	-

Table 7 displays the computation cost in terms of floating-point operations (FLOPs), model parameter count, as well as training and test time, of all studied band selection algorithms on the WHU-Hi-HanChuan dataset for selecting 60 bands. The training time is measured over five runs with different seeds. Notably, of all the deep-learning-based methods, WAST (Sokar et al., 2022) is the most light-weight and fast because it employs sparse training and a sparse autoencoder. SIFDR (Zhang et al., 2024) is a non-deep-learning-based statistical method and thus requires the least amount of time to generate band ranking.

Dropout feature ranking (FR) (Chang et al., 2017), stochastic gates (STG) (Yamada et al., 2020), and one-shot feature selection with additive noise distortion (SAND) (Pad et al., 2025) have the exact same architecture but differ in selection mechanisms. Interestingly, Dropout FR and STG both aim to relax discrete Bernoulli variables but utilize different distributions (Concrete vs. Gaussian) to approximate; both STG and SAND involve adding Gaussian noise to their input but do so differently with different constraints.

Of all deep-learning-based methods, SRL-SOA (Ahishali et al., 2022) is the only one that requires post-processing upon training to obtain band ranking, i.e., calculating the row sum of the representation matrix generated by the trained encoder. For others, we can directly obtain band ranking via sorting the learned model parameters (Dropout FR, STG, and SAND) or the data structure used to store neuron importance (WAST).

5.2.2 Phase 2: HSI Classification

Tables 8 and 9 showcase the computation cost and runtime for training and testing MambaHSI and SVM on the WHU-Hi-HanChuan dataset. In accordance with the findings in Li et al. (2024), spatial features are

Table 8: Comparison of computation cost on the WHU-Hi-HanChuan dataset for training MambaHSI with different numbers of bands

No. of Spectral Bands	FLOPs	Params	Training Time (s)	Test Time (s)	OA (%)
30 bands	5.13G	34.28K	35.77 ± 0.13	0.040	81.50 ± 1.95
60 bands	15.19G	101.75K	58.02 ± 0.41	0.066	89.15 ± 0.25
All Bands (SpaMB-ONLY)	26.70G	126.54K	46.71 ± 6.51	0.041	88.98 ± 2.49
All Bands (Full Model)	33.98G	137.01K	293.72 ± 29.93	0.311	89.87 ± 1.88

Table 9: Comparison of runtime on the WHU-Hi-HanChuan dataset for training SVM with different numbers of bands

No. of Spectral Bands	Training Time (s)	Test Time (s)	OA (%)
30 bands	0.01	5.93 ± 0.08	68.11 ± 0.61
60 bands	0.01	6.57 ± 0.16	69.28 ± 1.64
All Bands (272 bands)	0.02	17.38 ± 0.81	71.00 ± 0.90

more discerning than spectral features for HSI classification, and thus removing the Spectral Mamba Block only induce negligible accuracy degradation (∼1%) while reducing 84% of training time.

Unless specified with "Full Model", all experiments carried out with MambaHSI uses only the Spatial Mamba block (SpaMB). As noted in Appendix A, the hidden dimension of Mamba block decreases as the number of selected bands decreases. It can be observed from Table 8 that, when employing MambaHSI, using 30 bands ($\sim$ 11% of total bands) results in a mere $\sim$ 8% accuracy loss, and when using 60 bands ($\sim$ 22% of total bands), there's even a slight accuracy improvement ($\sim$ 0.15%). When employing SVM, the accuracy difference is rather small across a wide range of number of bands, as reflected by the relatively flat curves in Fig. 7 and Table 9. However, there's considerable savings in inference time when the number of bands is reduced.

5.3 Discussion

With respect to RQ #1, we determine that there doesn't exist a universally superior light-weight autoencoder-based HSI band selection method independent of the dataset and classifier in use. This is reasonable as there exists substantial diversity in each HSI dataset in terms of hyperspectral sensors used for data collection, spatial and spectral resolution, spectral range and scene type (e.g., urban or agricultural), etc.

Since it is common that HSI remote sensing datasets present low inter-class separability and high intra-class variability, SVM falls short since the classes most likely are not linearly separable. If a GPU-equipped edge device is available, MambaHSI will be a better option due to its light-weight architecture and fast inference. Also, a band selection method might not be necessary if MambaHSI is in use.

Regarding RQ #2, we learn that MambaHSI preserves accuracy fairly well – only a maximum 6% accuracy drop when using ∼20-30% of the bands (see Tables 3, 4 and 5). Table 8 also demonstrates the reduction in model FLOPs and parameters as well as runtime when using only a subset of bands.

The merit of this work is to analyze and compare the mechanism, architecture and computation cost for each method so that researchers, engineers and practitioners can make a more informed choice given their specific hyperspectral sensors, dataset, and constraints of computation resources.

6 Conclusions and Future Work

This work aligns two seemingly divergent research lines – the more general *feature selection* and the more specific *HSI band selection* and concentrates on efficient unsupervised deep learning-based techniques for

HSI band selection. We provide an in-depth comparative analysis of six state-of-the-art HSI band selection methods, including five light-weight autoencoder-based methods and one non-deep-learning method. Experiments are carried out on three publicly available remote sensing benchmark datasets with two classifiers – SVM and MambaHSI. The results show clear superiority of the Mamba-based model in terms of accuracy and inference time when a GPU is available. Although we could not conclude a universally superior band selection method that stands out across diverse datasets, we dived deep into the inner workings of each method, organized and put together the code in PyTorch, hoping it will benefit researchers and engineers in making the process of hyperspectral sensing and data processing more efficient and accurate. Future work will be to integrate one of the band selection methods into a real-world hyperspectral remote sensing and data processing pipeline with edge hardware onboard an unmanned aerial vehicle (UAV).

References

Arun Solomon A and Akila Agnes S. Land-cover classification with hyperspectral remote sensing image using cnn and spectral band selection. *Remote Sensing Applications: Society and Environment*, 31:100986, 2023. ISSN 2352-9385.

Mete Ahishali, Serkan Kiranyaz, Iftikhar Ahmad, and Moncef Gabbouj. SRL-SOA: Self-representation learning with sparse 1d-operational autoencoder for hyperspectral image band selection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2296–2300. IEEE, 2022.

Stefano B Blumberg, Paddy J Slator, and Daniel C Alexander. Experimental design for multi-channel imaging via task-driven feature selection. *arXiv preprint arXiv:2210.06891*, 2022.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Yaoming Cai, Xiaobo Liu, and Zhihua Cai. BS-Nets: An end-to-end framework for band selection of hyperspectral image. *IEEE Transactions on Geoscience and Remote Sensing*, 58(3):1969–1984, 2020.

Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers Electrical Engineering*, 40(1):16–28, 2014. ISSN 0045-7906. 40th-year commemorative issue.

Chein-I Chang, Qian Du, Tzu-Lung Sun, and M.L.G. Althouse. A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 37(6):2631–2641, 1999.

Chun-Hao Chang, Ladislav Rampasek, and Anna Goldenberg. Dropout feature ranking for deep learning models. *arXiv preprint arXiv:1712.08645*, 2017.

Weizhao Chen, Zhijing Yang, Jinchang Ren, Jiangzhong Cao, Nian Cai, Huimin Zhao, and Peter Yuen. MIMN-DPP: Maximum-information and minimum-noise determinantal point processes for unsupervised hyperspectral band selection. *Pattern Recognition*, 102:107213, 2020. ISSN 0031-3203.

Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J White, and Su-In Lee. Learning to maximize mutual information for dynamic feature selection. In *International Conference on Machine Learning*, pp. 6424–6447. PMLR, 2023.

Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.

Melika Dastranj, Timothy Smet, Courtney Wigdahl-Perry, Kenneth Chiu, Trevor Bihl, and Jayson Boubin. REMIX: Real-time hyperspectral anomaly detection for small UAVs. 05 2025.

Christian Debes, Andreas Merentitis, Roel Heremans, Jürgen Hahn, Nikolaos Frangiadakis, Tim van Kasteren, Wenzhi Liao, Rik Bellens, Aleksandra Pižurica, Sidharta Gautama, Wilfried Philips, Saurabh Prasad, Qian Du, and Fabio Pacifici. Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2405–2418, 2014.

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. ISSN 0031-3203.

Jie Feng, Jiantong Chen, Qigong Sun, Ronghua Shang, Xianghai Cao, Xiangrong Zhang, and Licheng Jiao. Convolutional neural network based on bandwise-independent convolution and hard thresholding for hyperspectral band selection. *IEEE Transactions on Cybernetics*, 51(9):4414–4428, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Isabelle M Guyon, Jason Weston, Stephen D. Barnhill, and Vladimir Naumovich Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

J. Ham, Yangchi Chen, M.M. Crawford, and J. Ghosh. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):492–501, 2005.

Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Sen Jia, Guihua Tang, Jiasong Zhu, and Qingquan Li. A novel ranking-based clustering approach for hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing*, 54(1):88–102, 2016.

Xiuping Jia and J.A. Richards. Fast k-nn classification using the cluster-space approach. *IEEE Geoscience and Remote Sensing Letters*, 2(2):225–228, 2005.

Ludmila I Kuncheva, Clare E Matthews, Álvar Arnaiz-González, and Juan J Rodríguez. Feature selection from high-dimensional data with very low sample size: A cautionary tale. *arXiv preprint arXiv:2008.12025*, 2020.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

Yapeng Li, Yong Luo, Lefei Zhang, Zengmao Wang, and Bo Du. MambaHSI: Spatial–spectral mamba for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Yufei Liu, Xiaorun Li, Ziqiang Hua, Chaoqun Xia, and Liaoying Zhao. A band selection method with masked convolutional autoencoder for hyperspectral image. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Shin-ichi Maeda. A bayesian encourages dropout. *arXiv preprint arXiv:1412.7003*, 2014.

F. Melgani and L. Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(8):1778–1790, 2004.

Fan Min, Qinghua Hu, and William Zhu. Feature selection with test cost constraint. *International Journal of Approximate Reasoning*, 55(1):167–179, 2014.

Giorgio Morales, John Sheppard, Riley Logan, and Joseph Shaw. Hyperspectral band selection for multispectral image classification with convolutional networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2021.

Pedram Pad, Hadi Hammoud, Mohamad Dia, Nadim Maamari, and L Andrea Dunbar. SAND: One-shot feature selection with additive noise distortion. *arXiv preprint arXiv:2505.03923*, 2025.

Billy G. Ram, Peter Oduor, C. Igathinathane, Kirk Howatt, and Xin Sun. A systematic review of hyperspectral imaging in precision agriculture: Analysis of its current state and future prospects. *Computers and Electronics in Agriculture*, 222:109037, 2024. ISSN 0168-1699.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Kevin Siebels, Kalifa Goïta, and Mickaël Germain. Estimation of mineral abundance from hyperspectral data using a new supervised neighbor-band ratio unmixing approach. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10):6754–6766, 2020.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Ghada Sokar, Zahra Atashgahi, Mykola Pechenizkiy, and Decebal Constantin Mocanu. Where to pay attention in sparse training for feature selection? *Advances in Neural Information Processing Systems*, 35: 1627–1642, 2022.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958, 2014.

He Sun, Lei Zhang, Lizhi Wang, and Hua Huang. Stochastic gate-based autoencoder for unsupervised hyperspectral band selection. *Pattern Recognition*, 132:108969, 2022. ISSN 0031-3203.

Weiwei Sun, Liangpei Zhang, Bo Du, Weiyue Li, and Yenming Mark Lai. Band selection using improved sparse subspace clustering for hyperspectral imagery classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):2784–2797, 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 12 1996. ISSN 0035-9246.

Madison Tuohy, Jasper Baur, Gabriel Steinberg, Jalissa Pirro, Taylor Mitchell, Alex Nikulin, John Frucci, and Timothy Smet. Utilizing UAV-based hyperspectral imaging to detect surficial explosive ordinance. *The Leading Edge*, 42:98–102, 02 2023.

Qi Wang, Fahong Zhang, and Xuelong Li. Optimal clustering framework for hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing*, 56(10):5910–5922, 2018.

Qi Wang, Fahong Zhang, and Xuelong Li. Hyperspectral band selection via optimal neighborhood reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8465–8476, 2020.

Qi Wang, Qiang Li, and Xuelong Li. A fast neighborhood grouping method for hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):5028–5039, 2021a.

Qian Wang, Jiaxing Zhang, Sen Song, and Zheng Zhang. Attentional neural network: Feature selection using cognitive feedback. *Advances in neural information processing systems*, 27, 2014.

Qian Wang, Li Sun, Yan Wang, Mei Zhou, Menghan Hu, Jiangang Chen, Ying Wen, and Qingli Li. Identification of melanoma from hyperspectral pathology image using 3d convolutional networks. *IEEE Transactions on Medical Imaging*, 40(1):218–227, 2021b.

Buyun Xu, Xihai Li, Weijun Hou, Yiting Wang, and Yiwei Wei. A similarity-based ranking method for hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing*, 59(11):9585–9599, 2021.

Lei Xu, Mete Ahishali, and Moncef Gabbouj. Dropout concrete autoencoder for band selection on hyperspectral image scenes. *IEEE Geoscience and Remote Sensing Letters*, 22:1–5, 2025.

Yutaro Yamada, Ofir Lindenbaum, Sahand Negahban, and Yuval Kluger. Feature selection using stochastic gates. In *International conference on machine learning*, pp. 10648–10659. PMLR, 2020.

Chenglong Zhang, Lichao Mou, Xiaoli Yang, Xiangrong Zheng, Xiao Xiang Zhu, and Xiaopeng Ma. A real-time unsupervised hyperspectral band selection via spatial-spectral information fusion-based downscaled region. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Yanfei Zhong, Xin Hu, Chang Luo, Xinyu Wang, Ji Zhao, and Liangpei Zhang. WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF. *Remote Sensing of Environment*, 250:112012, 2020. ISSN 0034-4257.

A Hyperparamter Details

In addition to the general hyperparameters for training a neural network, such as batch size, learning rate, and training epochs, each band selection method has its own hyperparameters to tune. Table 10 lists the number of hyperparameters uniquely for each band selection method to consider. We generally follow the insights and suggestions from the original papers, do a mini grid search on the Houston dataset, and then apply the same hyperparameters on the other two HSI datasets. To enhance the reproducibility of our work, the meaning of each hyperparameter and the values we use are provided below:

Table 10: Number of hyperparameters for each band selection algorithm

	SRL-SOA	Dropout FR	STG	WAST	SAND	SIFDR
# Hyperparam.	1	2	2	4	2	2

- SRL-SOA (Ahishali et al., 2022)
  - Q = {1, 3}; the number of 1D convolutional layers in the encoder.
- Dropout FR (Chang et al., 2017; Xu et al., 2025)
  - The start ($\tau_s = 1$) and final ($\tau_f = 0.01$) temperatures for the concrete selector. We apply an annealing scheduler (Eq. 3) such that the temperature gradually decays at each epoch. As the temperature τ gradually anneals to zero, the Concrete distribution approximates more closely to the discrete *argmax*.

$$\tau = \tau_s \times (\frac{\tau_f}{\tau_s})^{\frac{\text{curr\_epoch}}{\text{total\_epochs}}} \tag{3}$$

- STG (Yamada et al., 2020; Sun et al., 2022)
  - $\sigma = 0.01$; noise component of the stochastic gate is drawn from $\mathcal{N}(0, \sigma^2)$.
  - $\lambda = 0.1$; regularization coefficient.
- WAST (Sokar et al., 2022)
  - N_hidden = 200; number of neurons in hidden layer.
  - $\lambda = 0.4$; regularization coefficient in neuron importance.
  - $\alpha = 0.3$; the fraction of dropped and regrown weights.

  - density = 0.2; percentage of model weights/connections retained throughout training.

- SAND (Pad et al., 2025)
  - $\sigma = 1.5$; indicates how firmly we would like to restrict the number of features to k. A higher value of σ places greater emphasis on precisely achieving k features, resulting in faster binarization (polarization towards 0 and 1) of the gains (a_i's).
  - $\alpha = 2$; Euclidean norm is used to normalize the gain vector $\underline{a}$ during training.

- SIFDR (Zhang et al., 2024)
  - spatial scaling factor $r = \frac{1}{8}$ for bicubic interpolation to scale down the image.
  - batch size = 1000 for calculating the weight vector indicating band importance based on constrained energy.

Table 11 presents the general training hyperparameters for training deep-learning based band selection algorithms on all studied datasets.

For classifier training, we completely follow (Li et al., 2024) for training the Mamba classifier with a 0.0003 learning rate and 200 total epochs. The hidden dimension of the Mamba block is set to 64 when using all bands, and k when using k selected bands where $k < 64$. Note that when k is close to 64 (e.g., 60) as exhibited in Table 8, the runtime might slightly increase.

For the SVM classifier, we use the default `SVC` from `sklearn.svm` with the regularization parameter C set to 1000.

Table 11: General training hyperparameters for each deep-learning-based band selection algorithm

	SRL-SOA	Dropout FR	STG	WAST	SAND
batch size	15	15	15	5	15
learning rate	0.0001	0.0001	0.001	0.05	0.001
total epochs	50	50	50	10	50