Ingénierie des Systèmes d'Information
Vol. 31, No. 5, May, 2026, pp. 1495-1509
Journal homepage: http://iieta.org/journals/isi

Reconstruction-Guided Spatial–Spectral Feature Compression with a Convolutional Autoencoder–CNN for Hyperspectral Image Classification

Annisa Divayu Andriyani[1], Kamarul Hawari Ghazali[1,2*]

[1] Faculty of Electrical and Electronics Engineering Technology, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan 26600, Malaysia
[2] Center of Advanced Industrial Technology, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan 26600, Malaysia

Corresponding Author Email: kamarul@umpsa.edu.my

https://doi.org/10.18280/isi.310509

Received: 6 March 2026
Revised: 25 April 2026
Accepted: 5 May 2026
Available online: 31 May 2026

Keywords:
hyperspectral image classification, convolutional autoencoder, feature compression, convolutional neural network, reconstruction-guided learning, latent representation

ABSTRACT

Hyperspectral image (HSI) classification is challenged by high spectral dimensionality, redundancy among adjacent bands, and limited labelled samples. This study presents a reconstruction-guided spatial–spectral convolutional autoencoder–CNN (SSCAE–CNN) framework that couples compact feature learning with supervised classification. Spatial–spectral patches are first encoded into a latent representation by a convolutional autoencoder; the decoder regularizes the representation through patch reconstruction, while a CNN classifier refines the latent features for class prediction. The model was evaluated on Indian Pines, Salinas, and WHU-Hi HanChuan using the same stratified train–test protocol for all compared methods. Performance was assessed using overall accuracy (OA), average accuracy, Kappa, precision, recall, and F1-score, with five independent runs used to examine stability. Under the adopted experimental protocol, SSCAE–CNN achieved OA values of 99.42%, 99.67%, and 99.38% on Indian Pines, Salinas, and WHU-Hi HanChuan, respectively, and outperformed the evaluated SVM, Random Forest, 2D-CNN, 3D-CNN, CNN-GRU, and ViT baselines. Ablation experiments showed that the reconstruction pathway contributed to the quality of the learned representation, while latent-space visualizations and classification maps indicated improved class separation and spatial consistency. The results support reconstruction-guided feature compression as a useful strategy for HSI classification. Further work should assess the method under spatially disjoint splits and cross-scene transfer settings.

1. INTRODUCTION

From traditional machine learning methods to advanced ensemble techniques, numerous approaches have been developed for hyperspectral imagery classification [1]. In contrast to traditional RGB or multispectral imaging, which offer restricted spectral resolution, hyperspectral imaging (HSI) facilitates accurate characterization of material features through their distinct spectral signatures [2]. This rich spectral information allows for improved discrimination between different materials, even when they appear visually similar. Consequently, HSI has attained significant value across various applications, such as land cover classification, precision agriculture, environmental monitoring, mineral exploitation, and urban analysis [3].

Recent advancements in satellite-based hyperspectral sensors have significantly expanded the accessibility and applicability of HSI data. Spaceborne platforms such as PRISMA, EnMAP, and DESIS provide high-quality hyperspectral data for large-scale Earth observation with improved spatial and spectral resolution [4]. These technologies provide ongoing surveillance of environmental and surface conditions at both regional and global levels [5]. Consequently, they support critical applications such as ecosystem assessment, disaster monitoring, and natural resource management in a more efficient and scalable manner [6].

Despite its advantages, HSI data presents several challenges for analysis and classification tasks [7]. The elevated dimensionality of hyperspectral data, potentially encompassing hundreds of spectral bands, amplifies computational complexity and may impair classification perdormance due to the phenomenon known as the "curse of dimensionality" [8]. Furthermore, strong correlations among adjacent spectral bands introduce significant redundancy, which may hinder effective feature learning. In addition, the limited availability of labeled samples often leads to overfitting, especially when complex models are employed [9].

HSI classification methods are generally divided into conventional machine learning (ML) and deep learning (DL) approaches to tackle these difficulties [10]. Conventional machine learning techniques, including Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbours (KNN), predominantly depend on manually engineered features, resulting in a constrained ability to

encapsulate intricate spatial-spectral correlations [11]. Conversely, DL methodologies can autonomously acquire hierarchical feature representations directly from unprocessed data. Convolutional Neural Networks (CNN) and 3D-CNN architectures are extensively utilized for spatial-spectral feature extraction, whereas Recurrent Neural Networks (RNN), such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), are applied to represent spectrum relationships. Additionally, autoencoder-based approaches have been extensively explored for feature learning and dimensionality reduction [12].

However, existing ML and DL approaches still suffer from several limitations. Many methods attempt to improve classification accuracy by increasing model depth and architectural complexity, which often leads to higher computational cost and reduced efficiency [8]. Moreover, spectral redundancy and compact feature representation are not explicitly addressed in many existing frameworks. As a result, there remains a need for models that can effectively reduce dimensionality while preserving discriminative information for accurate classification [13].

This research introduces a Spatial–Spectral Convolutional Autoencoder and Convolutional Neural Network (SSCAE-CNN) framework to overcome these constraints in hyperspectral image classification. Unlike conventional autoencoder-based approaches that mainly use latent features for dimensionality reduction or reconstruction, the proposed framework explicitly connects spatial–spectral feature compression with CNN-based discriminative classification. The encoder is designed to learn compact latent representations from local hyperspectral patches, while the classifier directly refines these compressed features for final class prediction. The novelty of this study does not lie in merely combining an autoencoder with a CNN. Instead, it lies in the use of a reconstruction-guided latent space as a compact spatial–spectral representation for classification. This design aims to reduce spectral redundancy, preserve local spatial context, and improve classification accuracy with a relatively simple and efficient architecture.

This study's primary contributions are summarised as follows. A spatial-spectral convolutional autoencoder is developed to compress high-dimensional hyperspectral patches while maintaining critical spatial and spectral information. Second, the learned latent representation is directly integrated with a CNN classifier to improve discriminative feature learning. The proposed framework is assessed using three benchmark hyperspectral datasets and compared with conventional machine learning methods, deep learning models, and ablation variants to illustrate the efficacy of the proposed representation-learning technique.

2. RELATED WORK

Deep learning techniques have been widely employed for hyperspectral imaging classification due to their ability to independently extract hierarchical and discriminative features from high-dimensional datasets [14]. Among these techniques CNNs are among the most extensively used approaches because they can effectively extract spatial features from local image patches. 2D-CNNs generally treat spectral bands as input channels and focus on learning spatial patterns from hyperspectral patches [15]. In contrast, 3D-CNNs are designed to jointly capture spatial and spectral information through volumetric convolution operations [16]. Hybrid CNN architectures that combine 2D and 3D convolutional operations have also been proposed to exploit the advantages of both approaches and improve spatial–spectral feature representation [17].

To model spectral dependencies more effectively, RNNs have also been explored for HSI classification. In RNN-based approaches, spectral bands are commonly treated as sequential inputs, allowing the model to learn dependencies across wavelength dimensions [18]. Architectures such as LSTM and GRU are capable of capturing long-range spectral relationships and have demonstrated promising performance in hyperspectral classification tasks [19]. Recent studies have further integrated recurrent structures with convolutional layers to enhance joint spatial–spectral learning and improve classification accuracy [16].

Attention-based and transformer-based models have recently gained significant interestin hyperspectral image classification [20]. Attention mechanisms, including channel attention, spatial attention, and self-attention, improve feature learning by emphasizing informative spectral or spatial components while suppressing less relevant information [21]. Transformer-based designs enhance this ability by capturing long-range connections and global contextual linkages within the input data. These models provide an alternative to conventional convolutional and recurrent networks, particularly for learning complex spatial–spectral interactions in high-dimensional hyperspectral data [22].

Hybrid deep learning frameworks that combine multiple architectures have also shown improved performance in HSI classification. Examples include CNN-RNN, CNN-attention, CNN-transformer, and autoencoder-CNN models, which integrate feature extraction, sequence modeling, attention-based refinement, and representation learning within a unified framework [10]. These hybrid approaches are useful because HSI data contain both spatial structures and spectral dependencies that are difficult to capture using a single model type. However, many existing hybrid and autoencoder-based methods primarily focus on improving classification accuracy or reconstructing input data. They often do not explicitly emphasize compact spatial–spectral representation learning for reducing dimensional complexity while preserving discriminative information [23].

Autoencoder-based models are essential for addressing the high dimensionality and redundancy seen in hyperspectral data [24]. Autoencoders perform feature compression and representation learning by transforming high-dimensional input data into a lower-dimensional latent space. This procedure enables the model to retain critical information while minimising superfluous spectral elements. An autoencoder comprises two fundamental components: an encoder and a decoder. The encoder compresses the input into a latent representation, while the decoder reconstructs the original input from this compressed format [25].

Given an input sample $\boldsymbol{x} \in \mathbb{R}^d$, the encoder transforms the input into a latent representation as follows Eq. (1):

$$\boldsymbol{z} = f_\theta(\boldsymbol{x}) = \sigma_e(\boldsymbol{W}_e\boldsymbol{x} + \boldsymbol{b}_e) \tag{1}$$

where, $\boldsymbol{z} \in \mathbb{R}^k$, with $k < d$, represents the compressed latent feature vector [26]. The parameters $\boldsymbol{W}_e$ and $\boldsymbol{b}_e$ denote the encoder weight matrix and bias vector, respectively. The function $\sigma_e(\cdot)$ represents a nonlinear activation function used in the encoder. The decoder then reconstructs the input from

the latent representation as follows Eq. (2):

$$\hat{\boldsymbol{x}} = g_{\phi}(\boldsymbol{z}) = \sigma_d(\boldsymbol{W}_d\boldsymbol{z} + \boldsymbol{b}_d) \quad (2)$$

where, $\hat{\boldsymbol{x}}$ denotes the reconstructed input. The parameters $\boldsymbol{W}_d$ and $\boldsymbol{b}_d$ represent the decoder weight matrix and bias vector, respectively. The function $\sigma_d(\cdot)$ denotes the decoder activation function, which may differ from the encoder activation depending on the reconstruction objective and input normalization [27].

The purpose of the autoencoder's training is to decrease the reconstruction error between the original input and the reconstructed output. This objective is commonly formulated using the squared Euclidean distance as follows Eq. (3):

$$\mathscr{L}_{rec} = \| \boldsymbol{x} - \hat{\boldsymbol{x}} \|_2^2 \quad (3)$$

This loss function encourages the learned latent representation to preserve the most important information from the original input. By minimizing the reconstruction loss, the autoencoder learns compact features that reduce dimensionality while maintaining representative data characteristics [28]. In the context of HSI classification, this property is useful for reducing spectral redundancy and mitigating the computational burden caused by high-dimensional spectral bands.

To better capture spatial–spectral characteristics, convolutional autoencoders (CAEs) have been introduced by replacing fully connected transformations with convolutional operations [29]. Unlike conventional autoencoders, CAEs preserve local spatial structures while simultaneously learning correlations across spectral information. CAEs are especially appropriate for hyperspectral data, as both spatial context and spectral characteristics are crucial for precise classification. Recent research has shown that CAE-based frameworks can significantly boost feature compactness and improve classification performance in hyperspectral imaging classification tasks [30].

Although AE- and CAE-based methods have been widely used for hyperspectral image classification, several limitations remain. Many autoencoder-based approaches mainly focus on reconstruction quality or spectral dimensionality reduction, so the learned latent features are not always optimized for class discrimination. In addition, several CAE-CNN frameworks treat feature compression and classification as separate stages, which may weaken the interaction between representation learning and discriminative feature refinement. CNN-based models can extract local spatial features effectively, but direct classification from high-dimensional hyperspectral patches may still be affected by spectral redundancy and computational burden. Meanwhile, RNN- and transformer-based models can capture long-range dependencies, but they often require higher computational resources. Therefore, a more integrated framework is needed to connect reconstruction-guided spatial–spectral compression with supervised classification. In response to this gap, the proposed SSCAE-CNN uses the SSCAE encoder to learn compact latent representations from hyperspectral patches and directly feeds them into a CNN classifier, allowing classification to operate on compressed but informative spatial–spectral features.

3. METHODOLOGY

Current deep learning methods for hyperspectral image classification have demonstrated promising performance; yet, numerous systems continue to handle high-dimensional spectral data directly or regard feature compression and classification as distinct phases. This may reduce the effectiveness of the learned representation because the compressed features are not always aligned with the classification objective. This study offers a SSCAE-CNN architecture to tackle this issue, which combines reconstruction-guided feature compression with CNN-based discriminative learning in a unified framework.

The proposed framework is designed to learn compact latent representations from hyperspectral patches while preserving both spatial and spectral information. Unlike a conventional CAE-CNN pipeline, where the autoencoder is commonly used only as a preprocessing or dimensionality-reduction module, the proposed SSCAE-CNN uses the latent representation as the central feature space for classification. The decoder guides the encoder to preserve essential spatial–spectral information during training, while the CNN classifier refines the learned latent features for class prediction. This structure allows the model to reduce spectral redundancy while maintaining discriminative information for hyperspectral image classification.

To improve the understanding and reproducibility of the proposed framework, the key implementation details of SSCAE-CNN are summarised in Table 1. The proposed architecture consists of an encoder, a decoder, and a CNN classifier. The encoder compresses the input hyperspectral patch into a concise latent representation, while the decoder reconstructs the input during the training phase. The CNN classifier subsequently refines the latent features and produces the final class prediction through the Softmax layer.

Table 1. Detailed architecture of the proposed SSCAE-CNN framework

Component	Layer	Kernel Size	Feature Maps / Units	Activation	Output Description
Input	Hyperspectral patch	-	P × P × B	-	Spatial–spectral input patch
Encoder	Conv2D	3 × 3	32	ReLU	Low-level spatial–spectral features
Encoder	Conv2D	3 × 3	64	ReLU	Compressed feature representation
Encoder	MaxPooling2D	2 × 2	-	-	Reduced spatial dimension
Latent Space	Conv2D	3 × 3	64	ReLU	Compact latent representation
Decoder	UpSampling2D	2 × 2	-	-	Spatial feature restoration
Decoder	Conv2D	3 × 3	32	ReLU	Reconstructed feature refinement
Decoder	Conv2D	3 × 3	B	Sigmoid / Linear	Reconstructed hyperspectral patch
Classifier	Conv2D	3 × 3	64	ReLU	Discriminative feature extraction
Classifier	Global Average Pooling	-	-	-	Feature vector
Classifier	Dense	-	128	ReLU	Fully connected feature learning
Output	Dense	-	Number of classes	Softmax	Final class prediction

3.1 Overall framework

The proposed SSCAE-CNN framework has two primary components: an SSCAE and a CNN-based classifier. The SSCAE is tasked with feature compression and representation learning, while the CNN classifier executes discriminative feature learning and classification. The architecture allows the model to extract compact spatial-spectral featuress from hyperspectral data and directly use them for classification. This integration allows the latent representation to preserve meaningful information while supporting class discrimination.

The key distinction of the proposed framework is the role of the latent representation. In this study, the latent feature space is not treated only as a compressed version of the input, but as a shared representation that links reconstruction learning and classification. The reconstruction process encourages the encoder to retain essential spatial–spectral information, while the classifier encourages the learned features to be useful for distinguishing land-cover classes. Therefore, the proposed SSCAE-CNN provides a more task-oriented use of convolutional autoencoder features compared with standard AE-based representation learning.

Figure 1 depicts the comprehensive architecture of the proposed framework. The input hyperspectral image is first preprocessed and transformed into spatial–spectral patches. Each patch encompasses the spectral signature of the target pixel along with its adjacent spatial context, enabling the model to capture both spectral and local spatial information. This patch-based representation is important because hyperspectral classification depends not only on spectral characteristics but also on spatial relationships among neighboring pixels.

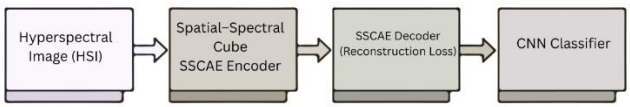

Figure 1. Block diagram of the proposed SSCAE-CNN framework

In the first stage, hyperspectral data are prepared by extracting spatial–spectral patches from the original hyperspectral cube. This stage offers structured input samples that maintain the correlation between spatial neighbourhoods and spectral bands, considering the high dimensionality of hyperspectral pictures. The extracted patches are then used as input to the proposed SSCAE module. This preparation step ensures that the model can learn features from both spectral signatures and spatial patterns.

In the second phase, the extracted patches are processed by the SSCAE module. The encoder compresses the high-dimensional input into a compact latent representation by identifying essential spatial-spectral characteristics. The decoder reconstructs the input from the latent representation to guarantee the preservation of crucial information during compression. This reconstruction process encourages the latent space to retain meaningful features while reducing redundancy in the original hyperspectral data.

In the conclusive phase, the acquired latent representation is transmitted to the CNN-based classifier for feature enhancement and classification. The CNN classifier additionally derives hierarchical and discriminative patterns from the compressed information. The refined features are subsequently converted to class probabilities by a fully linked layer and a Softmax function. Through this integrated design, the proposed framework performs both feature compression and classification in a unified manner.

The workflow of the proposed framework is further illustrated in Figure 2. The process begins with hyperspectral image input, followed by spatial patch extraction and spatial–spectral feature compression using the SSCAE. The model is trained by minimizing the reconstruction loss so that the encoder can learn compact and informative latent features. After training, the extracted latent features are used by the CNN classifier to perform final prediction.

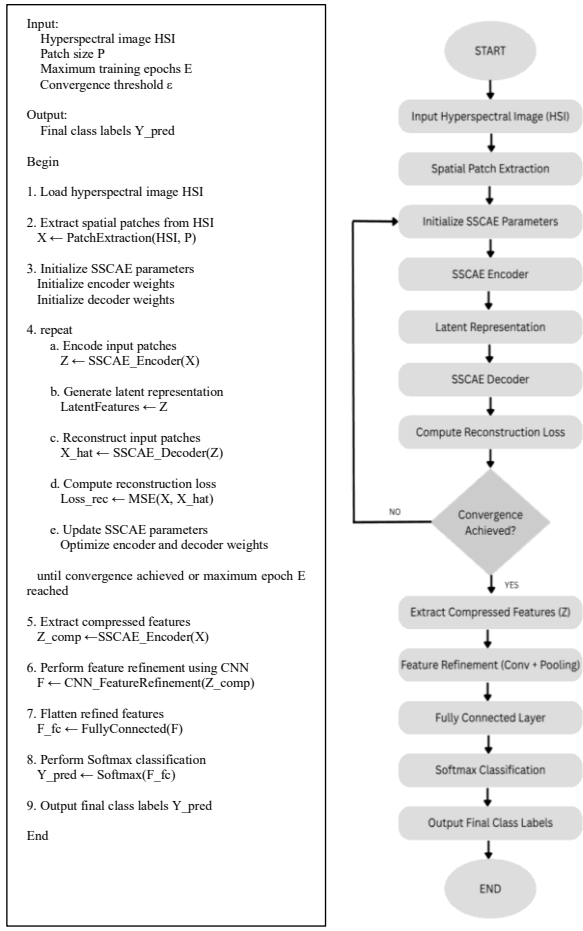

Figure 2. Flowchart and pseudocode of the proposed framework

3.2 Proposed Spatial-Spectral Convolutional Autoencoder

The proposed SSCAE is designed to learn compact and informative representations from hyperspectral image patches. Unlike conventional autoencoders that often operate on flattened spectral vectors, the proposed SSCAE processes three-dimensional spatial–spectral patches. This enables the encoder to capture local spatial patterns and spectral correlations simultaneously. The compressed latent representation is then used not only for reconstruction but also as the input feature space for the CNN classifier. This dual function makes the latent representation more relevant for classification than a purely reconstruction-oriented autoencoder.

The SSCAE utilises an encoder-latent-decoder architecture. The encoder transforms the input hyperspectral patch into a lower-dimensional latent representation, whilst the decoder

reconstructs the original input from this compressed format. The reconstruction objective guarantees that the latent space preserves critical spatial-spectral information from the input data. The compressed latent features are simultaneously forwarded to the CNN classifier to facilitate discriminative learning.

The overall architecture of the proposed model is illustrated in Figure 3. The SSCAE module performs feature compression and representation learning, while the CNN classifier performs feature refinement and classification. This architecture is designed to reduce spectral redundancy and computational complexity without losing important classification-related information. Therefore, the proposed framework can provide a balance between compact feature representation and classification performance.

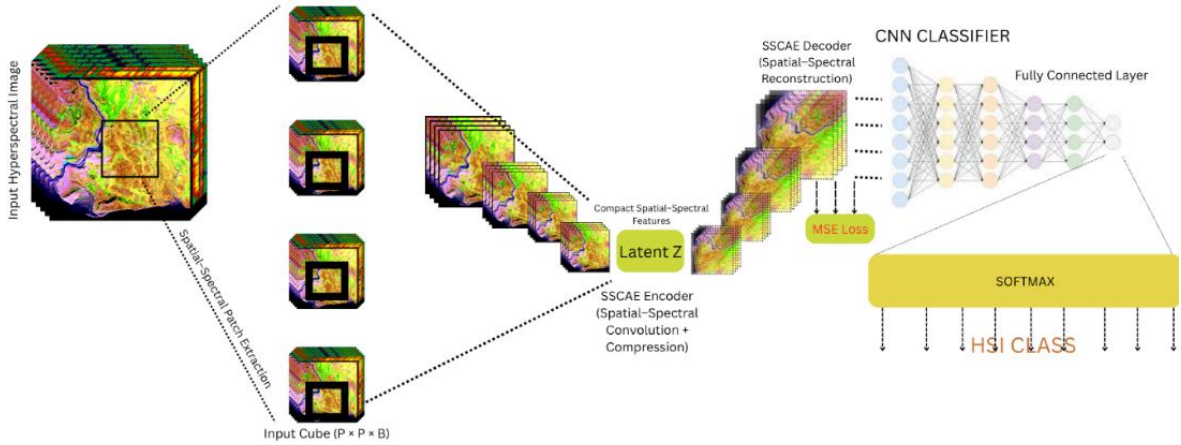

Figure 3. Comprehensive architecture of the proposed SSCAE-CNN framework for hyperspectral image classification

3.2.1 Encoder

The encoder component of the proposed SSCAE is designed to extract and compress spatial–spectral features from the input hyperspectral patch. Given an input hyperspectral patch $X \in \mathbb{R}^{P\times P\times B}$, P denotes the spatial patch size and B represents the number of spectral bands. The encoder processes the input using a series of spatial–spectral convolutional layers. These layers are employed to collect local spatial features and spectral correlations from the hyperspectral data.

The encoder consists of several convolutional blocks that progressively transform the input into higher-level feature representations. Each convolutional block extracts meaningful spatial–spectral patterns while reducing redundant information from the input data. Feature compression can be achieved using strided convolution or pooling operations, which reduce the resolution of the feature maps and generate a compact latent representation. This process helps reduce computational complexity and improves the efficiency of the learned features. Formally, the encoding process can be expressed as Eq. (4):

$$Z = f_\theta(X) \tag{4}$$

where, $f_\theta(\cdot)$ denotes the encoder mapping function parameterized by θ, and Z represents the resulting latent feature representation. The learned representation Z captures the most salient spatial–spectral characteristics of the input data in a compact form. This latent representation is subsequently used for both reconstruction by the decoder and classification by the CNN classifier. Therefore, the encoder plays a central role in transforming high-dimensional hyperspectral input into compact and informative features.

3.2.2 Decoder

The decoder element of the proposed SSCAE reconstructs the original hyperspectral input from the compressed latent representation. Given the latent representation Z, the decoder aims to reconstruct an approximation of the input cube, denoted as $\hat{X} \in \mathbb{R}^{P\times P\times B}$. This reconstruction process ensures that essential spatial and spectral information is preserved during feature compression. In contrast to the encoder, which diminishes the input representation, the decoder incrementally reconstructs the spatial-spectral structure of the input data.

The decoder consists of a sequence of spatial–spectral reconstruction layers. These layers may be implemented using transposed convolution or upsampling operations followed by convolutional layers. The reconstruction layers gradually recover the spatial and spectral resolution of the compressed features. By maintaining the three-dimensional configuration of hyperspectral data, the decoder may more efficiently retrieve both local spatial patterns and spectral characteristics.

To guide the reconstruction process, the SSCAE is trained using a reconstruction loss function. Since the input is represented as a hyperspectral cube, the reconstruction loss is formulated between the original input X and its reconstructed output $\hat{X}$. In this study, the mean squared error (MSE) reconstruction loss is defined as follows Eq. (5):

$$\mathscr{L}_{rec} = \frac{1}{N}\sum_{i=1}^{N} \| X_i - \hat{X}_i \|_2^2 \tag{5}$$

where, N denotes the number of training samples, X_i represents the original input hyperspectral patch, and $\hat{X}_i$ denotes the reconstructed output. The reconstruction loss measures the difference between the input and its reconstruction. Minimizing this loss encourages the encoder to learn latent features that preserve essential spatial–spectral information. This process also helps reduce redundant information while maintaining meaningful data characteristics for classification.

3.2.3 Latent feature representation

The latent feature representation Z is the output of the encoder and serves as a compact spatial–spectral embedding of the original hyperspectral input. As defined in Eq. (4), Z is obtained through the encoder mapping function $f_\theta(\cdot)$. This

representation can be interpreted as a nonlinear transformation of the original hyperspectral cube into a reduced-dimensional feature space. In comparison to the original input X, the latent representation is more succinct and less redundant.

The latent representation captures both local spatial structures and spectral correlations from the input hyperspectral patch. This property is important because hyperspectral image classification requires effective modeling of both spectral signatures and spatial context. By compressing the input into a compact latent space, the model can reduce dimensional complexity while preserving important discriminative information. Therefore, Z provides a more efficient representation for subsequent classification.

A key characteristic of the proposed framework is that the latent representation has a dual role. First, it is used by the decoder to reconstruct the original input, ensuring that the compressed features retain essential information. Second, it is directly forwarded to the CNN classifier, where it is used for discriminative feature refinement and classification. This dual function encourages the latent space to be both reconstructive and discriminative, which improves the effectiveness of the proposed SSCAE-CNN framework.

3.3 CNN classification and Softmax output

The classification component of the proposed framework performs discriminative learning using the compact latent representation obtained from the SSCAE encoder. Instead of using raw high-dimensional hyperspectral data, the classifier receives compressed spatial–spectral features as input. This design reduces redundancy and computational complexity while preserving important information for classification. As a result, the CNN classifier can focus on learning discriminative patterns from a more compact feature representation.

The latent features are analysed by a CNN classifier consisting of convolutional layers, nonlinear activation functions, pooling operations, and fully connected layers. The convolutional layers refine the latent representation by learning higher-level spatial–spectral patterns. Nonlinear activation functions, such as ReLU, improve the discriminative capability of the model by introducing nonlinearity into the learned features. Pooling operations further reduce feature dimensionality and improve robustness against small spatial variations.

After refinement, the resulting feature vector is transmitted through a fully connected layer to provide class scores or logits. The logits are subsequently converted into class probabilities through the application of the Softmax function. The Softmax function for class c is defined as follows in Eq. (6):

$$P(y = c \mid Z) = \frac{exp(z_c)}{\sum_{k=1}^{C} exp(z_k)} \tag{6}$$

where, C denotes the number of classes, and z_c represents the output score for class c. The class with the highest probability is selected as the final prediction.

By integrating the CNN classifier with the SSCAE framework, the proposed model effectively combines feature compression and discriminative learning. The latent representation ensures compact and informative features, while the CNN classifier enhances classification performance by learning complex feature relationships. This yields a resilient and effective model for hyperspectral image classification.

3.4 Training objective

The training process of the proposed framework is designed to optimize both feature reconstruction and classification performance. The SSCAE component is first trained to minimize the reconstruction loss so that the encoder can learn compact latent representations that preserve essential spatial–spectral information. The learned latent features are subsequently used as input for the CNN classifier in a supervised classification task. This training strategy ensures that the compressed representation remains meaningful and useful for classification tasks.

For the classification task, the CNN classifier is trained using the cross-entropy loss function. This loss quantifies the disparity between the expected class probability and the actual label. The classification loss is delineated as follows Eq. (7):

$$\mathscr{L}_{cls} = -\sum_{c=1}^{C} y_c \, log(P(y = c \mid Z)) \tag{7}$$

where, y_c denotes the ground truth label in one-hot encoded form, and $P(y = c \mid Z)$ represents the predicted probability for class c. The objective of the classifier is to minimize this loss by increasing the predicted probability of the correct class. This promotes the model's learning discriminative features that may successfully differentiate various land-cover categories.

The overall objective of the proposed framework can be formulated by combining reconstruction loss and classification loss. This joint objective encourages the model to learn latent features that are both reconstructive and discriminative. The total loss is defined as follows Eq. (8):

$$\mathscr{L}_{total} = \mathscr{L}_{cls} + \lambda \mathscr{L}_{rec} \tag{8}$$

where, λ is a weighting parameter that controls the contribution of the reconstruction loss. A larger value of λ gives more emphasis to reconstruction quality, while a smaller value gives more emphasis to classification performance. Through this combined objective, the proposed framework learns compact spatial–spectral features that preserve important information and support accurate classification.

3.5 Classification procedure

The classification procedure begins by extracting spatial–spectral patches from the hyperspectral image. Each patch is centered on a target pixel and contains both its spectral signature and surrounding spatial context. These patches are normalized and fed into the trained SSCAE encoder to generate compact latent representations. The resulting latent features are then used as input to the CNN classifier.

During inference, the decoder is not required for final classification because its main role is to support representation learning during training. The encoder produces latent features from the input patches, while the CNN classifier translates these features into class probabilities. The class exhibiting the highest Softmax probability is designated as the predicted label for the respective pixel. The procedure is repeated for all target pixels to generate the final classification map.

The proposed procedure allows the model to perform classification using compressed spatial–spectral features instead of raw hyperspectral data. This reduces computational burden while maintaining the discriminative information required for accurate classification. The proposed model utilises the encoder and classifier during inference, ensuring efficiency and practically for hyperspectral image classification applications. The final classification map illustrates the spatial distribution of predicted categories throughout the hyperspectral scene.

4. EXPERIMENTAL SETUP

4.1 Datasets

The performance of the proposed model was evaluated using widely used hyperspectral image datasets, which offer spatial–spectral information and a variety of land-cover categories. Each dataset consists of a three-dimensional cube with dimensions $H \times W \times B$, where H and W represent spatial dimensions and B denotes the number of spectral bands.

Prior to training, the hyperspectral images were preprocessed and converted into spatial–spectral patches using a sliding window approach. Specifically, patches of size $P \times P \times B$ (e.g., $9 \times 9 \times B$) were extracted around each labeled pixel to preserve local spatial context. These patches were then used as input samples for training and testing.

The datasets were divided into training and testing sets following standard protocols to ensure fair comparison with existing methods. The detailed characteristics of the datasets used in this study are summarized in Table 2.

Table 2. Description of hyperspectral datasets used in the experiments

Dataset Name	Spatial Size (H × W)	Spectral Bands (B)	Number of Classes	Training Samples	Testing Samples
Indian Pines [31]	145 × 145	200	16	6,149	2,050
Salinas [31]	512 × 217	224	16	32,477	10,826
WHU-Hi-HanChuan [32]	1217 × 303	274	16	154,518	51,506

4.2 Experimental protocol and reliability setting

To maintain an appropriate evaluation process, all models were trained and evaluated using the same dataset splitting approach. The labelled samples from each hyperspectral dataset were partitioned into training and testing sets by using a stratified sampling method to maintain the class distribution in both subsets. This setting was applied to ensure that each land-cover class was represented during model training and evaluation.

Since hyperspectral datasets may contain strong spatial correlation between neighboring pixels, the patch extraction process was performed after the training and testing indices were determined. Therefore, each target pixel was first assigned to either the training or testing subset, and the surrounding spatial context was then extracted as input for that target pixel. This procedure was used to reduce ambiguity in the evaluation protocol and minimize the risk of label leakage between training and testing samples.

In addition to the single-run comparison with baseline models, the proposed SSCAE-CNN was further evaluated using repeated independent runs with different random seeds. This additional experiment was conducted to assess the stability of the proposed model under different initialization and sampling conditions. The repeated-run results are reported using the mean, standard deviation, and boxplot visualization. This analysis provides additional evidence regarding the consistency of the proposed framework, although the baseline comparison remains based on the main experimental setting.

This study acknowledges that repeated-run analysis was mainly conducted for the proposed model. Therefore, future work may extend the repeated-run setting to all baseline models and include statistical significance testing for a more comprehensive robustness comparison.

4.3 Evaluation metrics

The classification performance of the proposed model was assessed using standard metrics in hyperspectral image classification, such as Overall Accuracy (OA), Average Accuracy (AA), Kappa coefficient, Precision, Recall, and F1-score. These measures evaluate overall performance, accuracy by class, and classification reliability.

OA quantifies the ratio of accurately identified samples to the total dataset, as defined in Eq. (9):

$$OA = \frac{\sum_{i=1}^{C} n_{ii}}{N} \times 100 \tag{9}$$

where, n_{ii} represents the number of correctly classified samples for class i, C denotes the total number of classes, and N represents the total number of samples. AA represents the mean classification accuracy across all classes and is calculated as Eq. (10):

$$AA = \frac{1}{C} \sum_{i=1}^{C} \frac{n_{ii}}{n_i} \times 100 \tag{10}$$

where, n_i represents the total count of samples in class i. The Kappa coefficient assesses the concordance between expected and actual labels. Precision quantifies the ratio of accurately predicted samples to the total predicted samples, whereas Recall assesses the ratio of correctly identified samples to the total real samples. The F1-score denotes the harmonic mean of Precision and Recall, offering a balanced assessment of classification performance.

In addition to the quantitative metrics, t-SNE visualization was applied to the learned latent features to provide qualitative insight into feature separability. The high-dimensional latent representations generated by the SSCAE encoder were projected into a two-dimensional space, allowing the class distribution and cluster separation to be visually analyzed. This analysis complements the quantitative metrics by illustrating how the proposed framework organizes hyperspectral samples in the latent feature space.

4.4 Repeated experiment protocol

To enhance the reliability of the experimental assessment,

the proposed SSCAE-CNN model was evaluated through multiple independent runs. For each dataset, training and testing were repeated over five independent runs using different random initializations. The same training and testing procedure was maintained across all iterations to guarantee uniformity. The results were then reported using the mean and standard deviation of the primary performance metrics, including OA, AA, Precision, Recall, F1-score, Kappa coefficient, training duration, inference duration, and GPU memory consumption. This repeated-run technique was adopted to evaluate the stability of the proposed model and to mitigate the likelihood that the stated results were obtained from a single favorable run.

In addition, fixed experimental settings were used during the repeated runs, including the same patch extraction strategy, model configuration, optimizer setting, batch size, and maximum number of epochs. Randomness was controlled by setting random seeds for the main computational libraries where applicable. The use of repeated runs provides a more reliable estimate of model performance and allows the robustness of the proposed framework to be analyzed through performance distribution.

5. RESULTS AND DISCUSSION

The performance of the proposed SSCAE-CNN framework was evaluated using three benchmark hyperspectral image datasets, namely Indian Pines, Salinas, and WHU-Hi HanChuan. The evaluation was conducted through quantitative performance assessment, repeated-run stability analysis, latent feature visualization, qualitative classification-map analysis, training convergence analysis, and computational time analysis. The quantitative evaluation includes OA, AA, Precision, Recall, F1-score, and Kappa coefficient. Repeated experiments were conducted to assess the stability of the proposed model, while t-SNE visualization was used to analyze the separability of the learned latent feature representations. In addition, classification maps, training curves, and time complexity results were used to provide further qualitative and computational insights into the model performance.

5.1 Quantitative classification performance using multiple evaluation metrics

The quantitative classification results on the Indian Pines, Salinas, and WHU-Hi HanChuan datasets are presented in Tables 3–5. The proposed SSCAE-CNN was compared with traditional machine learning methods, including SVM and Random Forest, as well as deep learning-based models, including 2D-CNN, 3D-CNN, CNN-GRU, and ViT. The inclusion of ViT provides a stronger comparison with a modern transformer-based baseline.

Overall, the proposed SSCAE-CNN achieved the highest classification performance across all three datasets. This result indicates that the integration of spatial–spectral feature compression and CNN-based discriminative classification improves the quality of learned representations for hyperspectral image classification. Compared with traditional machine learning methods, the proposed model produced substantially better results because it can learn spatial and spectral patterns directly from hyperspectral patches. In comparison to CNN-based and transformer-based baselines, the proposed model attained superior accuracy by reducing spectral redundancy via the SSCAE component prior to classification.

Table 3. Classification performance (%) on the Indian Pines dataset

Method	OA	AA	Kappa	Precision	Recall	F1-score
SVM	84.88	83.78	0.82	86.55	83.78	84.75
RF	79.17	71.33	0.75	80.03	71.33	74.18
2D-CNN	86.20	78.63	0.84	82.82	78.63	78.72
3D-CNN	93.66	92.23	0.92	95.42	92.23	93.57
CNN-GRU	96.34	86.30	0.95	88.66	86.30	86.85
ViT	95.37	95.20	0.94	94.61	95.20	94.76
SSCAE-CNN	**99.42**	**99.09**	**0.99**	**99.35**	**99.17**	**99.27**

Table 4. Classification performance (%) on the Salinas dataset

Method	OA	AA	Kappa	Precision	Recall	F1-score
SVM	86.84	88.97	0.85	90.92	88.97	89.49
RF	88.67	93.93	0.87	92.98	93.93	93.39
2D-CNN	87.85	81.58	0.86	82.69	81.58	81.13
3D-CNN	94.78	97.68	0.94	97.62	97.68	97.64
CNN-GRU	96.20	98.22	0.95	98.47	98.22	98.33
ViT	96.71	98.69	0.96	98.45	98.69	98.54
SSCAE-CNN	**99.67**	**99.85**	**0.99**	**99.83**	**99.85**	**99.84**

Table 5. Classification performance (%) on the WHU-Hi-HanChuan dataset

Method	OA	AA	Kappa	Precision	Recall	F1-score
SVM	84.88	83.78	0.82	86.55	83.78	84.75
RF	78.16	55.58	0.74	60.33	55.57	55.82
2D-CNN	88.19	66.27	0.86	75.59	66.27	67.06
3D-CNN	90.09	80.09	0.88	84.44	80.09	80.96
CNN-GRU	93.51	86.03	0.92	89.09	86.03	87.17
ViT	94.13	87.41	0.93	88.62	87.41	87.52
SSCAE-CNN	**99.38**	**98.90**	**0.99**	**98.96**	**98.90**	**98.93**

On the Indian Pines dataset, the proposed SSCAE-CNN obtained the best OA of 99.42%, AA of 99.09%, and F1-score of 99.27%. The improvement over 2D-CNN and 3D-CNN shows that direct convolutional classification may still be affected by spectral redundancy and limited feature compactness. Although CNN-GRU and ViT achieved competitive performance, the proposed model produced higher and more balanced results across all evaluation metrics. This confirms that reconstruction-guided spatial–spectral compression helps generate more discriminative latent features for classification.

For the Salinas dataset, SSCAE-CNN achieved the highest OA of 99.67%, AA of 99.85%, and F1-score of 99.84%. Salinas contains large homogeneous agricultural regions with clear spatial structures. This characteristic allows the proposed SSCAE module to effectively preserve spatial consistency while reducing redundant spectral information. The strong performance on Salinas indicates that the proposed framework is effective for agricultural hyperspectral scenes with structured land-cover distributions.

For the WHU-Hi HanChuan dataset, SSCAE-CNN achieved an OA of 99.38% and an F1-score of 98.93%. This dataset is larger and more complex than Indian Pines and Salinas, with more challenging spatial variations and class distributions. The lower AA and F1-score compared with Salinas indicate that some class-level variations remain more difficult to classify. However, the proposed model still outperformed all baseline models, demonstrating that compact spatial–spectral representation learning remains effective for larger UAV-borne hyperspectral scenes.

5.2 Repeated experiment and performance stability

To address the reliability of the reported high accuracy values, repeated experiments were conducted for the proposed SSCAE-CNN model. Each dataset was evaluated over five independent runs using the same experimental protocol, albeit with different random initializations. The findings are presented as mean ± standard deviation in Table 6. This repeated-run evaluation provides a more reliable estimate of model performance and reduces the possibility that the reported results are caused by a single favorable run.

The repeated-run results show that SSCAE-CNN produced stable performance across all datasets. On Indian Pines, the model achieved an OA of 99.42 ± 0.29%, indicating that the classification accuracy remained consistently high across repeated runs. On Salinas, the proposed model achieved an OA of 99.53 ± 0.37% and an F1-score of 99.79 ± 0.18%, showing stable performance on a high-resolution agricultural scene. On WHU-Hi HanChuan, the proposed model obtained an OA of 98.68 ± 0.16% and a Kappa coefficient of 0.9845 ± 0.0019. Although the class-wise metrics were slightly lower than those of Indian Pines and Salinas, the small standard deviation indicates that the model remained stable on a larger and more complex dataset.

Table 6. Repeated-run performance of the proposed SSCAE-CNN model on three hyperspectral datasets

Dataset	OA (%)	AA (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa	Training Time (s)	Inference Time (s)	GPU Memory (MB)
Indian Pines	99.42 ± 0.29	99.10 ± 0.98	99.48 ± 0.20	99.10 ± 0.98	99.27 ± 0.58	0.9934 ± 0.0033	35.0293 ± 5.2504	1.8308 ± 0.5267	2.1543 ± 0.1212
Salinas	99.53 ± 0.37	99.80 ± 0.14	99.78 ± 0.20	99.80 ± 0.14	99.79 ± 0.18	0.9948 ± 0.0041	215.6085 ± 90.1572	2.9697 ± 1.0474	2.2227 ± 0.1028
WHU-Hi HanChuan	98.68 ± 0.16	96.89 ± 0.62	97.60 ± 0.37	96.89 ± 0.62	97.20 ± 0.44	0.9845 ± 0.0019	430.4314 ± 140.6946	4.3868 ± 0.0890	2395.4742 ± 0.8846

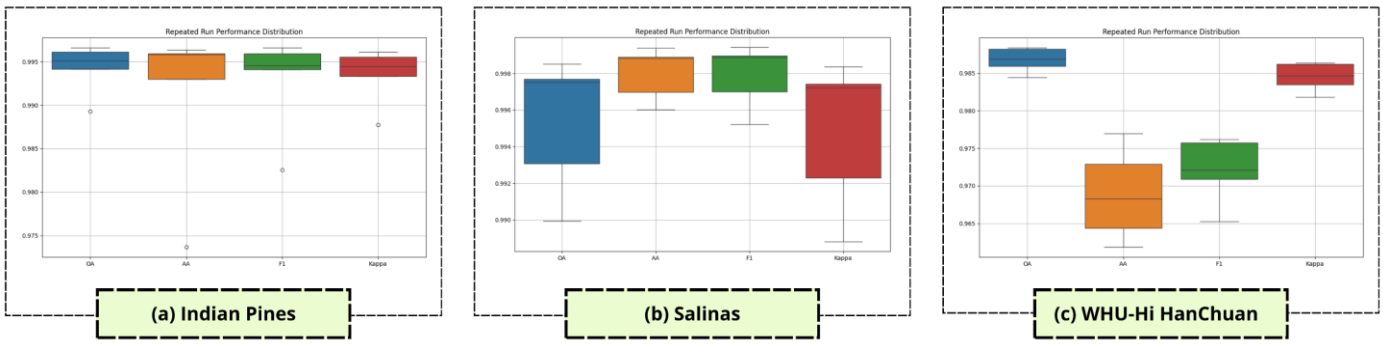

Figure 4. Repeated-run performance distribution of the proposed SSCAE-CNN model across datasets

The repeated-run performance distribution is shown in Figure 4. The boxplots indicate that OA and Kappa are concentrated within narrow ranges for all datasets. This suggests that the proposed model does not depend on a single favorable training run. The AA and F1-score values show slightly wider variation, particularly on WHU-Hi HanChuan, which may be related to class imbalance and more complex land-cover patterns. Nevertheless, the overall distribution confirms the stability of the proposed SSCAE-CNN framework.

5.3 Latent feature visualization using t-SNE

To further analyze the separability of the learned latent features, t-SNE visualization was applied to the feature representations generated by the proposed SSCAE-CNN model. The visualization was conducted on the Indian Pines, Salinas, and WHU-Hi HanChuan datasets, as shown in Figure 5. This analysis provides additional insights into how the proposed framework organizes hyperspectral samples in the latent feature space.

As shown in Figure 5, the Indian Pines and Salinas datasets show relatively clear class groupings, indicating that the proposed model can learn distinguishable latent representations for many land-cover classes. The Salinas dataset shows more clearly separated clusters, which is consistent with its high quantitative performance. In contrast, the WHU-Hi HanChuan dataset shows more overlapping regions among several classes. This suggests that the dataset contains more complex spatial–spectral variations and more challenging class boundaries. Nevertheless, several major clusters remain distinguishable, indicating that the learned representation still preserves useful class-related information.

Overall, the t-SNE visualization supports quantitative results by showing that the proposed SSCAE-CNN can form meaningful latent feature structures. However, the overlap observed in WHU-Hi HanChuan also indicates that further improvement may be needed for classes with similar spectral characteristics or complex spatial distributions.

5.4 Ablation study

An ablation study was conducted to evaluate the contribution of each component in the proposed SSCAE-CNN framework. Four model variants were compared: AE-only, CNN-only, SSCAE without decoder, and the full SSCAE-CNN model. The AE-only model evaluates the ability of reconstruction-based representation learning without a strong discriminative classifier. The CNN-only model evaluates direct classification without spatial–spectral feature compression. The SSCAE without decoder variant evaluates the effect of removing the reconstruction pathway. The full SSCAE-CNN model represents the complete framework with both feature compression and CNN-based classification.

The ablation results presented in Table 7 show that the full SSCAE-CNN consistently achieved the best performance across all datasets. The AE-only model produced lower performance because reconstruction-based feature learning alone does not guarantee class separability. Although the AE-only model can preserve general input information, its latent features may not be sufficiently discriminative for land-cover classification.

The CNN-only model achieved stronger performance than AE-only because it directly learns discriminative features through supervised classification. However, it still performed lower than the full SSCAE-CNN, indicating that direct CNN classification may still be affected by spectral redundancy and high-dimensional input representation. The SSCAE without decoder variant showed lower performance than the full model, confirming that the decoder and reconstruction objective play an important role during representation learning. The decoder helps guide the encoder to preserve essential spatial–spectral information, which improves the quality of the latent representation used by the classifier.

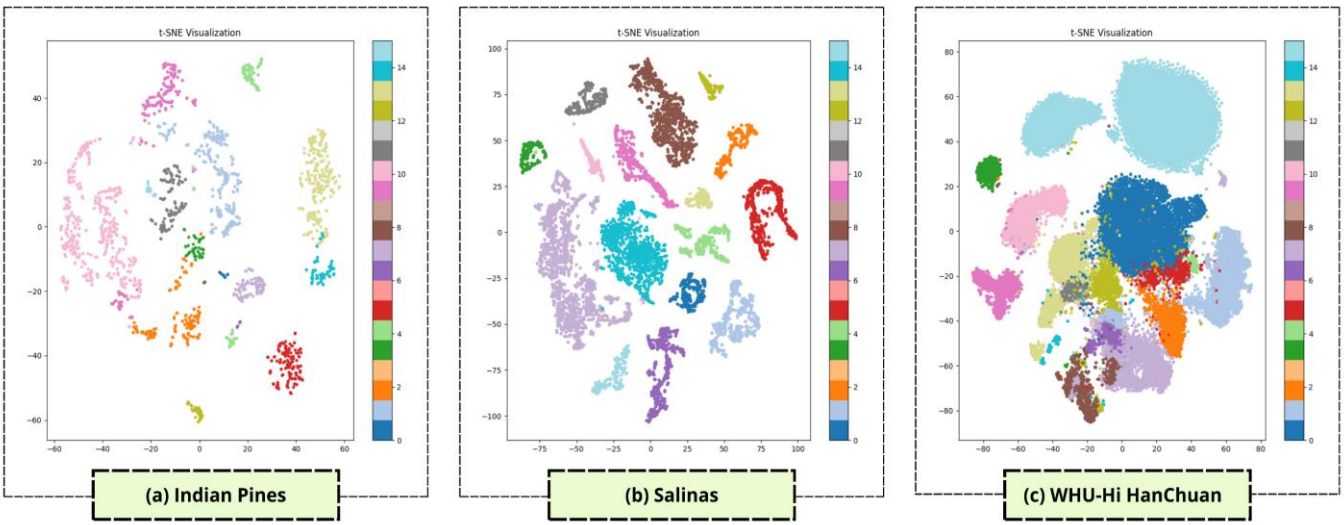

Figure 5. t-SNE visualization of latent feature representations learned by the proposed SSCAE-CNN model on three hyperspectral datasets: (a) Indian Pines, (b) Salinas, and (c) WHU-Hi HanChuan

Table 7. Ablation study on the across datasets

	Model Variant	OA	AA	Kappa	Precision	Recall	F1-score
Indian Pines	AE only	71.76	58.27	0.67	67.47	58.27	59.56
	CNN only	96.49	96.97	0.95	97.66	96.97	97.19
	SSCAE without decoder	59.80	55.91	0.55	64.27	55.91	51.49
	SSCAE-CNN	**99.27**	**99.17**	**0.99**	**99.35**	**99.17**	**99.26**
Salinas	AE only	86.46	91.19	0.84	92.71	91.19	91.19
	CNN only	96.34	98.36	0.95	98.21	98.36	98.26
	SSCAE without decoder	95.28	95.50	0.94	96.98	95.50	96.00
	SSCAE-CNN	**99.67**	**99.85**	**0.99**	**99.83**	**99.85**	**99.84**
WHU-Hi HanChuan	AE only	79.90	58.29	0.76	64.63	58.29	58.49
	CNN only	94.72	89.08	0.93	90.58	89.08	89.74
	SSCAE without decoder	95.57	90.33	0.94	91.39	90.33	90.69
	SSCAE-CNN	**99.38**	**98.90**	**0.99**	**98.96**	**98.90**	**98.93**

5.5 Classification maps

The qualitative classification maps are shown in Figures 6–8. These maps provide visual evidence of the spatial consistency and classification quality produced by each model. In general, traditional machine learning methods such as SVM and Random Forest produced more scattered misclassified pixels, especially in homogeneous regions. This indicates that traditional methods have limited ability to capture spatial–spectral relationships from hyperspectral data.

Deep learning models, including 2D-CNN, 3D-CNN, CNN-GRU, and ViT, produced smoother maps than SVM and Random Forest. However, some local noise and boundary errors were still observed, particularly in regions with spectrally similar classes. CNN-GRU and ViT produced competitive classification maps because they can model spectral dependencies and global feature interaction. Nevertheless, the proposed SSCAE-CNN generated maps that were visually closer to the ground truth.

For the Indian Pines dataset, the proposed model reduced scattered noise in agricultural regions and preserved class boundaries more effectively. For the Salinas dataset, SSCAE-CNN produced more homogeneous land-cover regions and fewer mixed pixels in large agricultural fields. For the WHU-Hi HanChuan dataset, the proposed model better preserved elongated field structures and reduced confusion along narrow boundaries. These qualitative results support the quantitative findings and show that the proposed framework improves not only numerical accuracy but also spatial coherence.

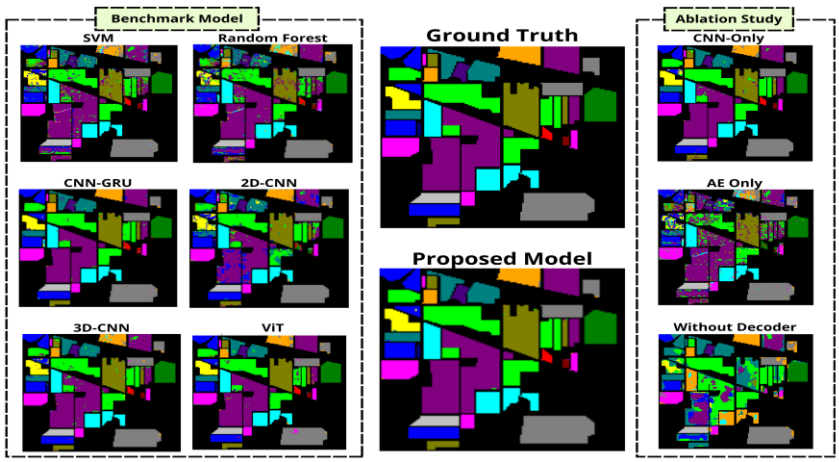

Figure 6. Classification maps of the Indian Pines dataset for different models

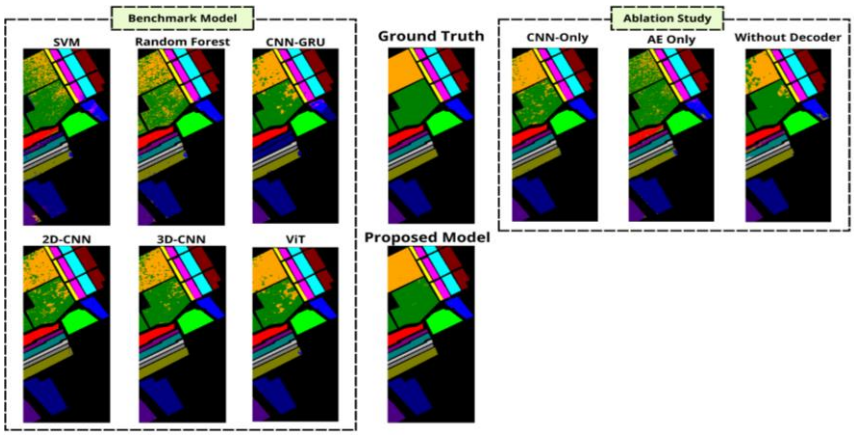

Figure 7. Classification maps of the Salinas dataset for different models

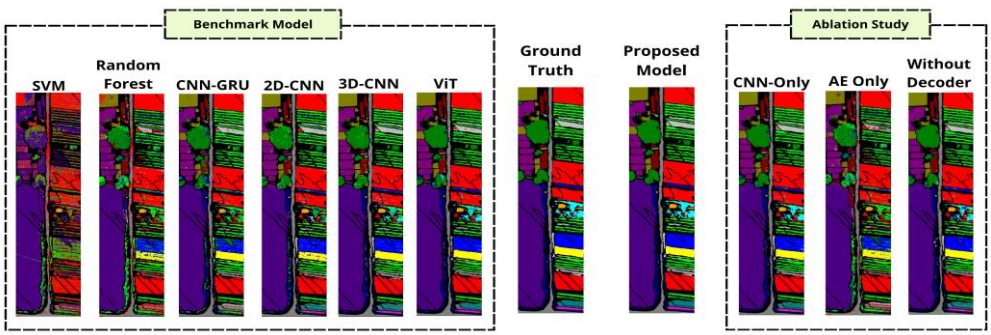

Figure 8. Classification maps of the WHU-Hi-HanChuan dataset for different models

5.6 Training convergence analysis

The training convergence behavior of the proposed SSCAE-CNN model is shown in Figure 9. The figure presents training and validation accuracy curves as well as classification loss curves for Indian Pines, Salinas, and WHU-Hi HanChuan. These curves help explain how the proposed model learns spatial–spectral representations during training.

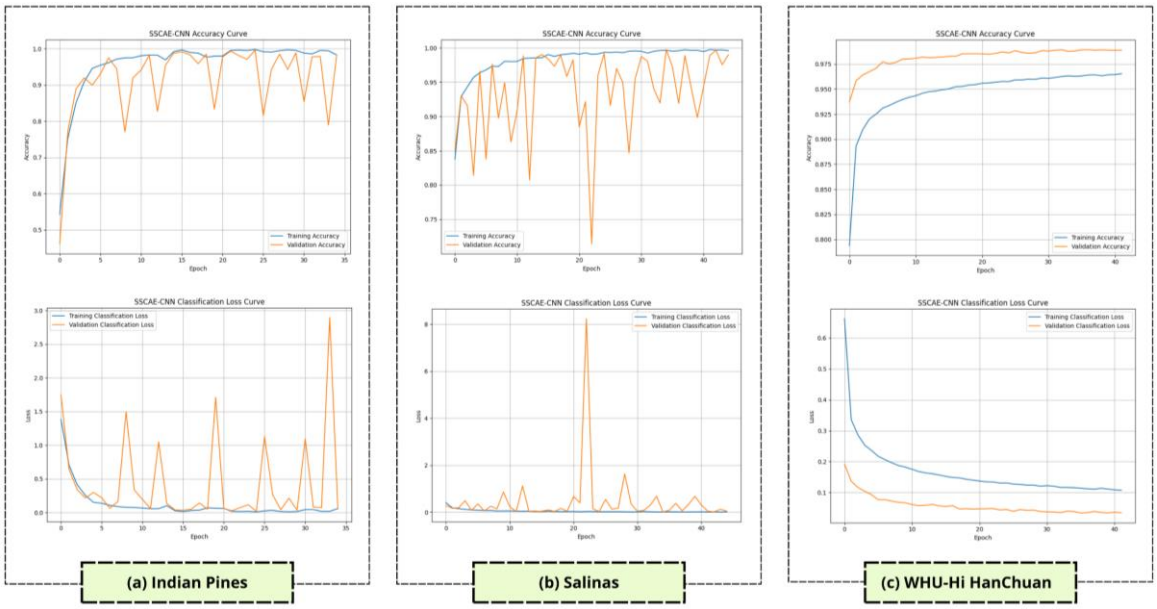

Figure 9. Training and validation accuracy and loss curves of the proposed model across datasets

For Indian Pines and Salinas, the training accuracy increased rapidly during the early epochs within a relatively small number of training epochs. However, the validation accuracy and validation loss showed several fluctuations. This behavior may be caused by the limited number of validation samples, class imbalance, and the sensitivity of small hyperspectral datasets to patch-based splitting. Therefore, repeated experiments are important to ensure that the final reported performance is not based on a single training condition.

For WHU-Hi HanChuan, the training and validation curves were smoother. Both training and validation accuracy improved gradually, while the loss decreased consistently. This indicates that the larger dataset provides more stable learning behavior. Overall, the convergence analysis shows that SSCAE-CNN can learn effective spatial–spectral representations across different dataset scales, although validation fluctuations on smaller datasets should be carefully considered.

5.7 Time complexity analysis

The computational complexity of the evaluated models was analyzed using training time and inference time. The results are presented in Table 8, while the visual comparisons are shown in Figure 10 and Figure 11. The evaluated models include 2D-CNN, 3D-CNN, CNN-GRU, ViT, and the proposed SSCAE-CNN.

The results show that 2D-CNN required the shortest training time due to its simple architecture. In contrast, ViT required the longest inference time across all datasets because transformer-based models generally involve more complex attention operations. The proposed SSCAE-CNN required longer training time than 2D-CNN, 3D-CNN, and CNN-GRU because the SSCAE component performs reconstruction-guided feature learning before classification. However, its inference time remained relatively efficient compared with ViT and CNN-GRU.

On WHU-Hi HanChuan, SSCAE-CNN required 1481.6441 s for training, which is higher than CNN-GRU and 3D-CNN but lower than ViT. More importantly, the inference time of SSCAE-CNN was only 1.1040 s, which is substantially lower than ViT and CNN-GRU. This indicates that the proposed model has a higher training cost but remains efficient during prediction. Consequently, SSCAE-CNN provides a favorable balance between classification accuracy and computational efficiency.

Table 8. Training time and inference time (in seconds) of different models across datasets

Dataset	Model	Training Time (s)	Inference Time (s)
Indian Pines	2D-CNN	8.2800	0.1965
	3D-CNN	26.3997	0.1778
	CNN-GRU	21.028	0.3752
	ViT	312.8243	2.4769
	SSCAE-CNN	93.6904	0.2354
Salinas	2D-CNN	33.5082	0.2762
	3D-CNN	62.3220	0.2831
	CNN-GRU	61.3594	0.6734
	ViT	448.1722	3.9243
	SSCAE-CNN	158.2657	0.3890
WHU-Hi-HanChuan	2D-CNN	227.9971	0.9682
	3D-CNN	704.1495	0.6655
	CNN-GRU	626.4376	2.9311
	ViT	1919.3671	11.0000
	SSCAE-CNN	1481.6441	1.1040

Overall, the time complexity analysis confirms that the proposed SSCAE-CNN achieves the best classification performance with moderate inference cost. Although the training process is more expensive due to the reconstruction-guided feature compression stage, the model remains practical for hyperspectral image classification because the decoder is mainly used during representation learning, while classification can be performed efficiently using the encoder and CNN classifier.

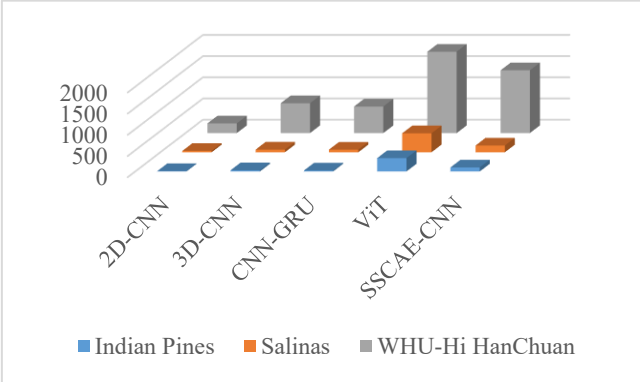

Figure 10. Comparison of training time (in second) for different models across the datasets

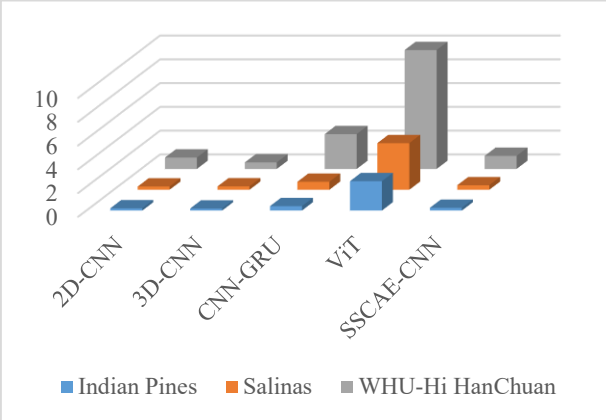

Figure 11. Comparison of inference time (in second) for different models across the datasets

6. CONCLUSIONS

This paper presented a SSCAE-CNN classifier for hyperspectral image classification. The proposed framework combines spatial–spectral feature compression with CNN-based classification to reduce spectral redundancy and learn compact latent representations from hyperspectral image patches.

Experiments on Indian Pines, Salinas, and WHU-Hi HanChuan showed that SSCAE-CNN achieved higher performance than the evaluated baseline models, including SVM, Random Forest, 2D-CNN, 3D-CNN, CNN-GRU, and ViT, under the experimental settings of this study. The repeated-run evaluation also showed relatively stable results across five independent runs, indicating that the reported performance was not obtained from a single training run.

The ablation study suggested that the combination of reconstruction-guided feature compression and CNN-based classification contributes to better classification results than AE-only, CNN-only, and SSCAE without decoder variants. The classification maps further indicated that the proposed model can produce more spatially consistent predictions with fewer scattered misclassified pixels in several regions.

However, this study still has several limitations. The evaluation was conducted on benchmark datasets using the defined train-test setting. No spatially disjoint validation, cross-dataset transfer experiment, or statistical significance test was included. Therefore, future work should further examine the model under stricter validation protocols, optimize the training cost, and explore lightweight attention or transformer-based extensions for broader hyperspectral image classification tasks.

ACKNOWLEDGMENT

This study received funding from the Ministry of Higher Education Malaysia through the Fundamental Research Grant Scheme (FRGS) (Grant No. FRGS/1/2024/TK07/UMP/01/1) and Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) (Grant No. RDU240107). The authors express their gratitude to MOHE and UMPSA for their financial support, research facilities, and ongoing assistance during the conduct of this work.

REFERENCES

[1] Dahiya, N., Singh, S., Gupta, S. (2023). A review on deep learning classifier for hyperspectral imaging. International Journal of Image and Graphics, 23(4): 2350036. https://doi.org/10.1142/S0219467823500365

[2] López-Baldomero, A.B., Martínez-Domingo, M.Á., Valero, E.M. (2026). Hyperspectral imaging for material identification in cultural heritage: A critical review. Applied Spectroscopy Reviews, 1-38. https://doi.org/10.1080/05704928.2026.2651333

[3] Srivastava, A., Jain, S. (2025). Remote sensing for environment assessment: Multispectral, hyperspectral, and thermal imaging applications. In Remote Sensing for Environmental Monitoring, pp. 1-31. https://doi.org/10.1007/978-981-96-5546-5_1

[4] Gámez García, J.A., Lazzeri, G., Tapete, D. (2025). Airborne and spaceborne hyperspectral remote sensing in urban areas: Methods, applications, and trends. Remote Sensing, 17(17): 3126. https://doi.org/10.3390/rs17173126

[5] Bourriz, M., Hajji, H., Laamrani, A., Elbouanani, N., Abdelali, H.A., François, B., Ali, E., Abdelhakim, A., Abdelghani, C. (2025). Integration of hyperspectral imaging and AI techniques for crop type mapping: Present status, trends, and challenges. Remote Sensing, 17(9): 1574. https://doi.org/10.3390/rs17091574

[6] Ilamathi, P., Chidambaram, S. (2025). Integration of hyperspectral imaging and deep learning for sustainable mangrove management and sustainable development goals assessment. Wetlands, 45(1): 9. https://doi.org/10.1007/s13157-024-01887-4

[7] Datta, D., Mallick, P.K., Bhoi, A.K., Ijaz, M.F., Shafi, J., Choi, J. (2022). Hyperspectral image classification: Potentials, challenges, and future directions. Computational Intelligence and Neuroscience, 2022(1): 3854635. https://doi.org/10.1155/2022/3854635

[8] Ahmad, M., Shabbir, S., Roy, S.K., Hong, D., Wu, X., Yao, J., Khan, A.M., Mazzara, M., Distefano, S., Chanussot, J. (2022). Hyperspectral image classification—Traditional to deep models: A survey for future prospects. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15: 968-999. https://doi.org/10.1109/JSTARS.2021.3133021

[9] Yang, J.Q., Wu, C., Du, B., Zhang, L.P. (2021). Enhanced multiscale feature fusion network for HSI classification. IEEE Transactions on Geoscience and Remote Sensing, 59(12): 10328-10347. https://doi.org/10.1109/TGRS.2020.3046757

[10] Ullah, F., Ullah, I., Khan, R.U., Khan, S., Khan, K., Pau, G. (2024). Conventional to deep ensemble methods for hyperspectral image classification: A comprehensive survey. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 3878-3916. https://doi.org/10.1109/JSTARS.2024.3353551

[11] Zaka, M.M., Samat, A. (2024). Advances in remote sensing and machine learning methods for invasive plants study: A comprehensive review. Remote Sensing, 16(20): 3781. https://doi.org/10.3390/rs16203781

[12] Zheng, Z.Y., Zhang, S.Z., Song, H.L., Yan, Q. (2024). Deep clustering using 3D attention convolutional autoencoder for hyperspectral image analysis. Scientific Reports, 14(1): 4209. https://doi.org/10.1038/s41598-024-54547-2

[13] Grewal, R., Kasana, S.S., Kasana, G. (2023). Machine learning and deep learning techniques for spectral spatial classification of hyperspectral images: A comprehensive survey. Electronics, 12(3): 488. https://doi.org/10.3390/electronics12030488

[14] Islam, M.R., Islam, M.T., Uddin, M.P., Ulhaq, A. (2024). Improving hyperspectral image classification with compact multi-branch deep learning. Remote Sensing, 16(12): 2069. https://doi.org/10.3390/rs16122069

[15] Aslam, M.A., Ali, M.T., Nawaz, S., Shahzadi, S., Fazal, M.A. (2023). Classification of rethinking hyperspectral images using 2D and 3D CNN with channel and spatial attention: A review. Journal of Engineering Research and Sciences, 2(4): 22-32. https://doi.org/10.55708/js0204003

[16] Ghazali, K.H., Andriyani, A.D., Yan, S.Q. (2025). Spectral spatial fusion using 3D CNN and attention-based Bi-LSTM for hyperspectral image classification. Journal of Applied Remote Sensing, 19(4): 046513. https://doi.org/10.1117/1.JRS.19.046513

[17] Mohamed Ali, M.S., Islam, M.S.B., Majid, M.E., Kashem, S.B.A., Khandakar, A., Chowdhury, M.E.H. (2025). A hybrid 3D CNN-LSTM model with soft spatial attention mechanism for accurate hyperspectral image classification. Remote Sensing Applications: Society and Environment, 40: 101779. https://doi.org/10.1016/j.rsase.2025.101779

[18] Gündüz, A., Orman, Z. (2025). Hyperspectral image classification using a hybrid RNN-CNN with enhanced attention mechanisms. Journal of the Indian Society of Remote Sensing, 53(2): 613-629. https://doi.org/10.1007/s12524-024-02011-z

[19] Arain, B., Ali, A.M., Alrashdi, I., Sallam, K.M., Abdel-Basset, M. (2025). Deep learning framework for land cover and land use classification: Five case studies with hyperspectral and RGB imagery. Neural Computing and Applications, 37(32): 26765-26822. https://doi.org/10.1007/s00521-025-11644-1

[20] Yang, J.H., Li, A.Q., Qian, J., Qin, J., Wang, L.G. (2024). A cross-attention-based multi-information fusion transformer for hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 13358-13375. https://doi.org/10.1109/JSTARS.2024.3429492

[21] Chen, J.D., Li, W.Z., El-Askary, H. (2025). Spectral-spatial transformer with multiscale convolutional attention for hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18: 24223-24241. https://doi.org/10.1109/JSTARS.2025.3608699

[22] Aleissaee, A.A., Kumar, A., Anwer, R.M., Khan, S., Cholakkal, H., Xia, G.S., Khan, F.S. (2023). Transformers in remote sensing: A survey. Remote Sensing, 15(7): 1860. https://doi.org/10.3390/rs15071860

[23] Liu, J.J., Wu, Z.B., Xiao, L., Wu, X.J. (2022). Model inspired autoencoder for unsupervised hyperspectral image super-resolution. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-12. https://doi.org/10.1109/TGRS.2022.3143156

[24] Gao, L.R., Li, J.X., Zheng, K., Jia, X.P. (2023). Enhanced autoencoders with attention-embedded degradation learning for unsupervised hyperspectral image super-resolution. IEEE Transactions on Geoscience and Remote Sensing, 61: 1-17. https://doi.org/10.1109/TGRS.2023.3267890

[25] Song, W.H., Zhang, X., Yang, G.Z., Chen, Y.J., Wang, L.C., Xu, H.H. (2024). A study on dimensionality reduction and parameters for hyperspectral imagery based on manifold learning. Sensors, 24(7): 2089. https://doi.org/10.3390/s24072089

[26] Sellami, A., Tabbone, S. (2022). Deep neural networks-based relevant latent representation learning for hyperspectral image classification. Pattern Recognition, 121: 108224. https://doi.org/10.1016/j.patcog.2021.108224

[27] Wu, H.J., Zhang, K.F., Wu, S.Q., Shi, S.S., Bian, C.F., Zhang, M.H. (2023). Unsupervised encoder-decoder network under spatial and spectral guidance for hyperspectral and multispectral image fusion. IEEE Transactions on Geoscience and Remote Sensing, 61: 1-16. https://doi.org/10.1109/TGRS.2023.3320404

[28] Kuester, J., Gross, W., Schreiner, S., Middelmann, W., Heizmann, M. (2023). Adaptive two-stage multisensor convolutional autoencoder model for lossy compression of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, 61: 1-22. https://doi.org/10.1109/TGRS.2023.3328222

[29] Lin, J.Y., Gao, F., Shi, X.C., Dong, J.Y., Du, Q. (2023). SS-MAE: Spatial-spectral masked autoencoder for multisource remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 61: 1-14. https://doi.org/10.1109/TGRS.2023.3331717

[30] Yu, W.B., Huang, H., Shen, G.X. (2023). Deep spectral-spatial feature fusion-based multiscale adaptable attention network for hyperspectral feature extraction. IEEE Transactions on Instrumentation and Measurement, 72: 1-13. https://doi.org/10.1109/TIM.2022.3222480

[31] Ayerdi, B., Graña Romay, M. (2016). Hyperspectral

image analysis by spectral-spatial processing and anticipative hybrid extreme rotation forest classification. IEEE Transactions on Geoscience and Remote Sensing, 54(5): 2627-2639. https://doi.org/10.1109/TGRS.2015.2503886

[32] Zhong, Y.F., Hu, X., Luo, C., Wang, X.Y., Zhao, J., Zhang, L.P. (2020). 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. https://doi.org/10.1016/j.rse.2020.112012

NOMENCLATURE

AA	Average Accuracy, dimensionless
B	Number of spectral bands in the hyperspectral image, dimensionless
C	Number of land-cover classes, dimensionless
$F1$	F1-score, dimensionless
H	Height of the hyperspectral image, pixel
$Kappa$	Kappa coefficient, dimensionless
N	Total number of samples, dimensionless
OA	Overall Accuracy, dimensionless
P	Spatial patch size, pixel
$P(y=c\|Z)$	Predicted probability that latent representation (Z) belongs to class (c), dimensionless
W	Width of the hyperspectral image, pixel
X	Input hyperspectral patch or cube
$\hat{X}$	Reconstructed hyperspectral patch or cube
Z	Latent spatial–spectral feature representation
c	Class index, dimensionless
$f_\theta(\cdot)$	Encoder mapping function
n_i	Total number of samples in class (i), dimensionless
n_{ii}	Number of correctly classified samples for class (i), dimensionless
z_c	Output score or logit for class (c), dimensionless
z_k	Output score or logit for class (k), dimensionless

Greek symbols

θ	Learnable parameters of the encoder
λ	Weighting parameter for reconstruction loss, dimensionless

Subscripts

c	Class index
I	Sample or class index
k	Class index used in Softmax summation
rec	Reconstruction
cls	Classification
$total$	Total objective function