Indian Journal of Pure & Applied Physics
Vol. 64, September 2026, pp. 1008-1017
DOI: 10.56042/ijpap.v64i9.31193

Multi-Scale Hybrid Spatial-Spectral Transformer for Hyperspectral Image Classification

Jay Kishor Sah*, Dipak Kumar Ghosh & Usha Chauhan

Department of Electrical Electronics and Communication Engineering, Galgotias University, Greater Noida 203 201, Uttar Pradesh, India

Received: 5th May 2026; accepted: 7th July 2026

Hyperspectral image classification (HSIC) requires the effective integration of spectral and spatial information to achieve accurate land-cover mapping. Conventional convolutional neural network (CNN) based methods are limited in modeling long-range dependencies and multi-scale contextual features inherent in hyperspectral data. In this research paper, a Multi-Scale Hybrid Spatial–Spectral Transformer (MSH-SST) framework for robust HSIC was proposed. The proposed architecture combines convolutional layers with transformer-based self-attention mechanisms to jointly learn local spatial patterns and global spectral-spatial relationships. Multi-scale feature extraction modules are employed to capture contextual information at different spatial resolutions, enhancing the representation of complex structures and boundary regions. A hybrid spatial spectral token embedding strategy is designed to preserve discriminative spectral information while maintaining spatial coherence. The transformer encoder further models long-range dependencies across both spectral and spatial dimensions, improving class separability. Experimental results conducted on well-known Indian Pines, Pavia University, and Salinas hyperspectral datasets validate that MSH-SST consistently outperformed over state-of-the-art CNN and transformer-based approaches in terms of classification accuracy and robustness, particularly under limited training sample conditions. The proposed framework effectively balances fine-grained spatial feature learning and global spectral-spatial dependency modeling, making it a powerful solution for high-precision hyperspectral image analysis.

Keywords: Deep learning, Hyperspectral image classification, Multi-scale learning, Spatial-spectral features, Transformer

1 Introduction

Hyperspectral imaging (HSI) is a technique for obtaining rich spectral information, where hundreds of contiguous spectral bands are captured for every pixel in an image, thus enabling accurate material identification in various remote sensing applications, such as land cover mapping, environmental monitoring, agriculture, and urban studies. Despite the availability of rich spectral information in hyperspectral images, hyperspectral image classification (HSIC) is a difficult task owing to the high dimensionality of hyperspectral data, spectral variability, mixed pixels, and lack of labeled samples[1]. Therefore, for effective hyperspectral image classification, it is important to exploit both spectral information and spatial contextual information to accurately classify complex surface materials.

In the past, various hyperspectral image classification (HSIC) approaches were proposed, mainly relying on spectral-based HSIC, such as support vector machines, k-nearest neighbors, and random forest classifiers. Although these approaches were found to perform well, they did not consider spatial information, thus yielding noisy classification maps for hyperspectral images. To overcome these limitations, spatial-spectral feature extraction approaches were proposed, such as morphological profiles, Gabor filters, and extended attribute profiles, thus yielding better classification accuracy for hyperspectral images[2]. Recently, deep learning approaches, mainly convolutional neural networks (CNNs), have been found to dominate the HSIC literature owing to the powerful feature learning capability of CNNs for hyperspectral image classification.

In CNNs, one-dimensional CNNs emphasize spectral features, two-dimensional CNNs emphasize spatial features, and three-dimensional CNNs emphasize both spectral and spatial features for hyperspectral image classification. Although CNNs have been found to perform well in hyperspectral

*Corresponding author: E-mail: jaykishor.sah@galgotiasuniversity.edu.in

image classification, they have been found to be limited in capturing long-range spatial dependencies in hyperspectral images, owing to the fixed convolutional kernels in CNNs, thus failing to emphasize discriminative features at different scales[3]. Recently, transformer models based on self-attention mechanisms have been shown to achieve remarkable performance on computer vision tasks by considering global dependencies and long-range contextual relationships[4]. Several transformer-based models have been proposed to solve HSIC problems and have been shown to achieve better performance than traditional CNN models. However, transformer models usually require large amounts of training data and may not consider fine local spatial structures that are very important for precise boundary and texture recognition in HSIs. In this paper, a novel Multi-Scale Hybrid Spatial-Spectral Transformer (MSH-SST) model is proposed to solve HSIC problems. The proposed model consists of convolutional neural networks to learn local spatial structures and transformer-based self-attention mechanisms to learn global spectral-spatial dependencies. A multi-scale feature extraction mechanism is used to learn contextual information. A hybrid spatial-spectral token embedding mechanism is proposed to learn discriminative spectral information while maintaining spatial coherence.

2 Literature Review

In the case of HSIC, the paradigm transition from classical learning strategies to advanced deep learning techniques enables modeling intricate dependencies between spectral and spatial dimensions. Initially, HSIC was based on simple spectral learning using conventional ML algorithms, such as support vector machine (SVM), k-nearest neighbor (KNN), random forest (RF), and sparse representation-based learning. Although this technique utilized the vast number of high-dimensional spectral data points effectively, it failed to consider spatial relationships among neighboring pixels, producing noisy results and failing to provide spatial consistency in output classes[5].

In response to these deficiencies, spatial-spectral techniques have been proposed, leveraging hand-crafted feature extraction, for instance, morphological profiles (MP), extended morphological attribute profiles (EMAP), Gabor filters, and local binary patterns (LBP). Though effective for improving classification performance, the utilization of manual feature extraction limited the model's expressiveness and generality across various hyperspectral datasets. With the advent of deep learning techniques, convolutional neural networks (CNNs) have emerged as the state-of-the-art solution for HSIC because of their ability to perform hierarchical feature representation learning[6]. In this respect, 3D CNNs proved especially efficient, learning spectral-spatial dependencies from raw data cubes to identify local correlations among adjacent pixels within the spectral and spatial domains. Yet, despite its merits, 3D CNN remains inherently restricted by localized receptive fields, limiting long-range dependency modeling. Moreover, predefined convolutional filters do not provide sufficient flexibility to learn adaptive features and degrade significantly in terms of performance in the presence of sparse training data[7].

As an alternative solution, transformers have been incorporated into HSIC through self-attention mechanisms to model global spectral-spatial correlations, dynamically representing the interaction between all pixels in the input feature space. Transformers constitute a powerful tool for encoding long-range contextual information; yet, they generally demonstrate low efficiency in capturing fine-grained local spatial structures, besides demanding significant amounts of annotated data and being computationally expensive.

Inspired by the complementary capabilities of CNNs and transformers, recent advances have focused on hybrid architectures combining convolutional layers with attention mechanisms. Within this context, this study proposes a hybrid 3D CNN-Transformer approach for HSIC. In the proposed architecture, the 3D CNN component serves as a local feature learner, extracting rich spectral-spatial representations from raw hyperspectral cubes. Transformers are then utilized to learn global dependencies and interactions among extracted features. To enhance spectral-spatial feature representation, multi-scale convolutional operations are applied inside the 3D CNN backbone to extract features at multiple scales. Such multi-scale representation increases the discriminative ability of the model in separating complex classes and preserving object boundaries[8]. By integrating multi-scale 3D convolutional feature learning with attention mechanisms, the proposed framework concurrently learns local spatial information, long-range contextual dependencies, and spectral-spatial representations. Consequently, the developed HSIC model demonstrates improved accuracy and robustness

under conditions of insufficient labeled samples and high intra-class variations.

3 Methodology

The proposed Multi-Scale Hybrid Spatial-Spectral Transformer (MSH-SST) framework for HSIC. The proposed approach combines CNNs with transformer-based attention mechanisms for joint handling of local spatial-spectral features and global spectral-spatial dependencies for HSIC. The proposed framework has four key components for data preprocessing, multi-scale spatial feature extraction, hybrid spatial-spectral token embedding, and transformer-based feature modeling with classification.

3.1 Data Preprocessing

HSI images are represented as three-dimensional data cubes with 2 spatial dimensions and 1 spectral dimension. Before training the proposed model, noisy bands and water bands are eliminated to reduce spectral redundancy. Additionally, each pixel value is normalized to zero mean and unit variance for enhanced model performance. To enable spatial-spectral feature extraction, fixed-size 3D patches are used to extract data patches centered at each target pixel in the HSI data cube.

3.2 Multi-Scale Spatial Feature Extraction

For the extraction of spatial structures at multiple scales, a multi-scale convolutional module is used. Multiple parallel branches with different sizes of convolutional kernels are used for the extraction of spatial features from input patches. Small-sized kernels are used for the extraction of fine textures and edges, while larger-sized kernels are used for the extraction of broader patterns. The feature maps from all branches are concatenated and fused using a 1x1 convolution for dimensionality reduction and improvement of feature consistency.

3.3 Hybrid Spatial–Spectral Token Embedding

After the extraction of spatial features, the feature maps from the previous stage are reshaped into a sequence of tokens, which are processed by the transformer encoder for feature modeling. In this stage, each token carries spatial as well as spectral information, retaining the discriminative spectral signatures of hyperspectral data while preserving spatial consistency. Positional embedding's are used for retaining spatial location information, which is critical for accurate land cover classification.

3.4 Transformer-Based Feature Modeling

In this stage, the embedded tokens are input into the transformer encoder, which comprises multiple blocks of multi-head self-attention (MHSA) and feed-forward networks (FFNNs). In this stage, self-attention is used for capturing long-range dependencies between spatial regions and spectral bands by assigning dynamic weights to the tokens. Layer normalization and residual connection are used for stabilization and improvement of training convergence.

3.5 Classification Layer

The output features produced by the transformer encoder are combined by a global pooling operation to generate a compact feature representation for the input patch. The compact feature representation produced by the global pooling operation is then passed through a fully connected layer followed by a SoftMax function to generate class probability scores for the input patch. The model is trained using a cross-entropy loss function, and the Adam optimizer with an adaptive learning rate is used for optimization.

3.6 Training Strategy

To resolve the issue of a small number of labeled samples, which is common in hyperspectral images, the training strategy adopted by the authors includes the application of data augmentation techniques like random flipping and rotation during training. Early stopping and dropout are implemented to prevent the model from overfitting during training. The model is trained end-to-end, which allows all the components to be trained for the best classification performance.

3.7 Patch Extraction and Embedding

In the patching process, the hyperspectral image is divided into local image patches, each of which contains the local information around a pixel, to create the input samples for the model. Each patch is assigned a ground truth label, which is utilized during the training process. Formally, the whole patching approach could be defined as a systematically well-ordered set of spectral-spatial patches that are being extracted from the hyperspectral image. In other words, the whole set of patches could be considered as a two-dimensional matrix, wherein each particular cell represents one patch placed at a definite location in the hyperspectral image. Thus, from the formal perspective, the patches matrix could be presented through the following notation: Patches (patch (i, j)),

where i equals 1, 2,…, n; j equals 1, 2,…, m. In this, the values of n and m represent the number of patches along the horizontal and vertical axes, respectively. As such, the patching operation allows one to transform the initial hyperspectral data cube into a matrix that represents local spectral-spatial information. Finally, the patches matrix obtained through patching serves as a basic input in the course of further feature extraction and learning processes such as the Transformer-based one.

The efficacy of the proposed MSH-SST framework has been validated in various applications such as detection, classification, and segmentation. Vision Transformers (VTs) have been successful in various traditional computer vision applications. Nevertheless, VTs are faced with some challenges in hyperspectral image classification. For instance, VTs are based on self-attention mechanisms that are insufficient in capturing local information in images. In HSI images, it is important to capture local information in order to achieve effective classification. Nevertheless, VTs are unable to capture local information as effectively as traditional CNN-based approaches. VTs are resolution-dependent and are unable to generalize effectively in HSI images of varying resolutions.

In order to address these limitations of VTs in HSI classification, an enhanced VT-based framework is proposed in this study that can effectively learn local and global information in HSI patches. The proposed framework can effectively classify HSI images by achieving highest classification accuracy and efficiency. The proposed architecture is based on the hybrid approach of cross-attention mechanisms, 3D convolutional layers, and graph attention network (GAN) layers that are further enhanced by the channel attention mechanism. This hybrid approach can effectively represent features in HSI images and can effectively classify all images. In the proposed framework, the HSI image is divided into various patches that are further extracted from the original image. Patch-based processing has several benefits for HSI analysis. Firstly, it allows for the inclusion of local spatial information, as the patches include information from the spatial neighborhood and the spectral information, which is important for HSI analysis. Secondly, it allows for localized feature learning, as features are learned independently from small regions in the image. Thirdly, it improves the model's efficiency by utilizing 3D convolutional neural networks on the patches, as they are able to capture effective spatial information and discriminative spectral-spatial features. In addition, the model incorporates more information from the HSI by adding positional encoding, as the features from the CNN are augmented and the positional value is added for each patch in the image. The hybrid block is significant for the development of a hierarchical structure for HSI features, which is important for effective HSI analysis as it allows for the inclusion of global and spatial dependencies. This is achieved through the inclusion of cross-attention mechanisms, which allow for the inclusion of global contextual patterns from the HSI image. Additionally, 3D-CNN GAT layers are included for the development of spatial relationships among the patches in the image. In these layers, the patches participate in an attention mechanism and assign weights to their relevance.

The attention mechanism is one of the fundamental theories in the field of deep learning, which enables a model to focus on those parts of the input data that are most significant while processing information. It differs from other techniques in that it does not assign an equal weightage to all features in the input but rather distributes weightage according to the significance of individual components. Through this technique, the model learns to identify significant features while ignoring or suppressing less significant ones.

3.8 3D-CNN GAT Layer and Cross-Attention Fusing

The cross-attention mechanism is integrated into the network by using a cross-attention operation that seamlessly combines information from both the cross-attention mechanism and the 3D-CNN GAT layers. The fusion of this information is crucial for obtaining a comprehensive interpretation of the HSI. The cross-attention mechanism is focused on the most informative patches within a specific region of the image, whereas the 3D-CNN GAT layers are responsible for modeling spatial dependencies between patches. The fused features are then fed into a series of dense layers during the global feature representation stage.

The classification stage of the MSH-SST model involves a dense layer and a SoftMax activation function. This activation function maps the features to probability distributions over classes. The probability distribution is a probability distribution over all possible classes. The MSH-SST model is trained on a labeled HSI dataset by minimizing the cross-entropy loss between the predicted probability distribution and

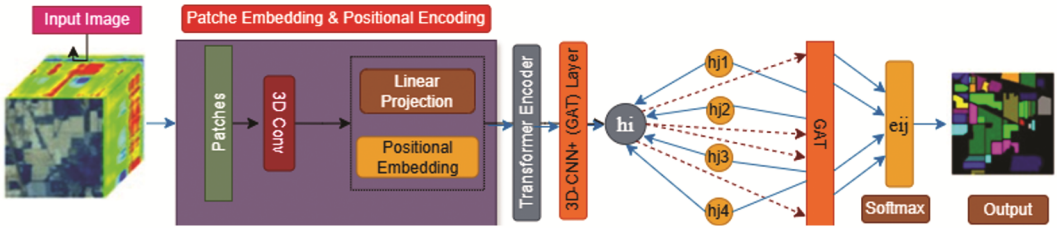


Fig. 1 — Architecture of proposed model for multi-scale hybrid spatial–spectral transformer

ground truth labels. The parameters are optimized using the Adam optimizer with a suitable learning rate. The OA, AA, and Kappa coefficients are used as evaluation metrics for the performance of the MSH-SST model. The architecture of the proposed MSH-SST model is presented in Fig. 1.

4 Proposed Model Results and Dataset Details

4.1 Proposed Model

All hyperspectral images are treated using a uniform experimental protocol. In this setting, each image is divided into three parts, namely 5 % training, 5 % validation, and 90 % testing datasets. Such division is carefully designed in order to evaluate the efficiency of the proposed model in cases where labelled data are limited, which is a common problem in HSI classification. SOTA models use the same data partitioning scheme as the proposed one. Baseline methods include 2D CNN, Hybrid CNN, Fast and Compact 3D CNN (FC3DCNN), 2D Inception Network (2D IN), 3D Inception Network (3D IN), and Hybrid Inception Network (Hybrid IN). Proposed method consists of an innovative hybrid architecture combining 3D-CNN and Transformer encoder. Proposed model is capable of capturing both local spectral-spatial correlations and long-range global contextual relations. Processing starts from inputting HSI cubes through 3D convolutional layers having a gradually growing number of filters. Various convolutional kernels are used to process raw hyperspectral data. This procedure allows extracting low and middle level spectral-spatial features, preserving structural properties of hyperspectral data. Batch normalization and ReLu activation are used after every convolution layer in order to accelerate convergence and improve feature learning stability.

The following stage consists of transforming output features to the token sequence and feeding them to Transformer encoder. Transformer encoder is based on multi-head self-attention (MHSA) block in order to capture long-range dependencies both in spectral and spatial domains. Position-wise feed forward networks (FFN) are additionally used in the attention block to enhance learning performance and avoid vanishing gradients problems with residual connection and layer normalization.

In order to increase performance by means of improving features representation, a lightweight fusion approach is used to combine local features obtained by 3D CNN and global representation obtained by Transformer. Features after the fusion are further processed by densely connected layers including dropout regularization (with rate = 0.4). The SoftMax classifier generates predictions for each of possible HSI classes.

Baseline approaches are compared by performance with the proposed architecture in terms of the metrics mentioned below. For 2D CNN, there are four convolutional layers (filters = 8, 16, 32, 64; kernel size = 3×3) before dense layers (filters = 256 and 128). FC3DCNN uses four 3D convolutional layers with different kernel sizes: 3×3×7, 3×3×5, and 3×3×3. Hybrid CNN involves 3D and 2D convolutions, transitioning from volumetric to planar representations. In a similar fashion, Inception models involve multi-branched 2D and 3D convolutional blocks of various sizes.

Performance of all methods is measured via standard metrics, namely Overall Accuracy (OA), Average Accuracy (AA), and Kappa (k). Overall accuracy (OA) is calculated as the percentage of correctly classified samples. In turn, average accuracy (AA) is the mean of per-class accuracies to balance performance in each class. The Kappa statistic measures an agreement between predictions and ground truth by correcting OA for the chance-level classification accuracy. Together, these metrics demonstrate the superior performance of the proposed architecture in capturing both local and global feature representations.

4.3 Experimental Dataset

In the current study, our analysis is based on well-known HSI datasets available in public sources, such

as remote sensing repositories. All images used in this research are acquired via airborne or satellite sensors, representing genuine Earth observation data. Moreover, the chosen datasets are annotated with reliable ground truth labels. A comprehensive experiment involving thorough evaluations of our HSI classification algorithm is conducted on each HSI dataset.

4.3.1 Indian Pines Dataset

The Indian Pines scene was acquired by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensor and covers agriculture fields in northwest Indiana, United States. The image has 145 by 145 pixels in terms of spatial dimension, consists of 224 spectral bands and wavelengths in the range between 0.4 and 2.5 × 10^{-6}. The majority of land cover in the scene is agricultural land; the other parts are forest, vegetation, roads, and buildings. Additionally, the image is recorded when early stages of crops are observed, for instance, corn and soybean with significant variation of spectra. The Indian Pines scene is categorized by ground truth labels into 16 different land cover classes. In order to remove water absorption spectral bands, spectral bands from 104 to 220 are eliminated[9]. Thus, the resulting dataset is represented with 200 bands shown in Fig. 2 (a).

4.3.2 Salinas Dataset

The Salinas dataset was created using the AVIRIS sensor and covers the Salinas Valley, California, United States. The scene has 224 bands in the spectral domain, 3.7 m/pix in spatial resolution and 512 by 217 pixels in the spatial dimension. As the result of pre-processing, 20 water absorption bands are deleted in order to improve signal quality[10]. The dataset provides information about the radiance of the scene and includes a variety of agricultural land cover types (vegetables, vineyards, bare soil). The scene is annotated by 16 classes of ground truth shown in Fig. 2 (b).

4.3.3 Pavia Centre Dataset

The Pavia Center scene was acquired using the Reflective Optics System Imaging Spectrometer (ROSIS) sensor and covers the urban environment in Pavia, Italy. The dataset has 102 spectral bands, 1.3 m/pixel spatial resolution, and 1096 × 1096 pixels. This image contains a mixture of land covers and illustrates a complex urban scenario[11]. The image is annotated by 9 different classes shown in Fig. 2 (c).

4.4 Implementation Details

Our proposed model is developed in Python utilizing an open-source libraries TensorFlow and Keras deep learning frameworks. TensorFlow is designed for algebraic computations with data flow graphs, while Keras provides an accessible interface for working with TensorFlow. The overall objective is to showcase the efficiency of the proposed model using a GPU. As a result, a constant batch size of 32 and learning rate of 0.0001 are used to evaluate all other models, including the proposed model. However, identifying the optimized parameters associated with the proposed model & SOTA models. These are crucial steps towards performance and optimization improvement.

Two parameters play a crucial role in model optimization. The first parameter is the percentage of training data samples that are necessary to learn the model properly. The second parameter is related to selecting an optimal patch size that is necessary to optimize HSI data. Initially, HSIs are divided into smaller overlapping patches. Using smaller patches in deep models is a crucial step towards optimization. Using smaller patches in deep models results in a significant reduction in time and parameters related to processing the all three datasets in a single iteration.

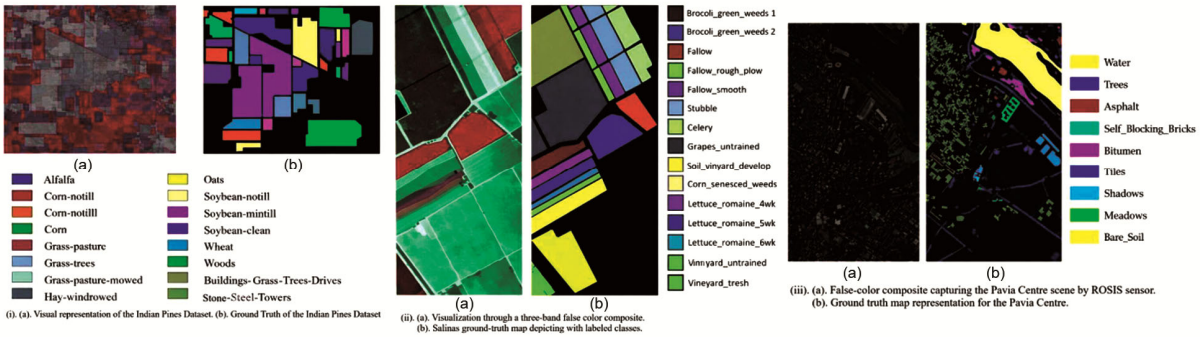


Fig. 2 — Experimental evaluation for all three hyperspectral datasets

As a result, splitting the HSI cube to the smaller patches is a crucial step towards optimization.

For efficient patching, firstly, all the cubes which contain the HSI and have sizes lower than the size of the patches should be padded by zeros. The patches should be labeled by their true class labels, and the next step would be the verification of the absence of the empty cubes. After that, we can continue the extraction of the training, validation, and testing samples from the patches. In particular, it is important to keep the precise ratio of the training samples and to provide the precise number and location of the samples to allow the comparison of our Proposed Model with the State-of-the-Art Models.

During the training, there is an implementation of the random initialization of the weights followed by optimization through backpropagation in the 3D CNN using Adam and SoftMax activation function in the last layer. All parameters used in the training of the model are fixed for all models. The results of training the SOTA and Proposed models are shown in the Fig. 3.

5 Results and Discussion

An in-depth analysis for each class for the proposed 3D CNN–Transformer is provided compared with other state-of-the-art (SOTA) methods. Table 1 presents the classification results for Indian Pines. Each row in this table represents one distinct land-cover (LC) class, while each column denotes a model that includes 2D Inception Network (2D IN)[12], 2D CNN[13], 3D Inception Network (3D IN)[14], Hybrid CNN[15], Hybrid Inception Network (Hybrid IN)[16], Fast and Compact 3D CNN (FC3DCNN)[17], and our proposed

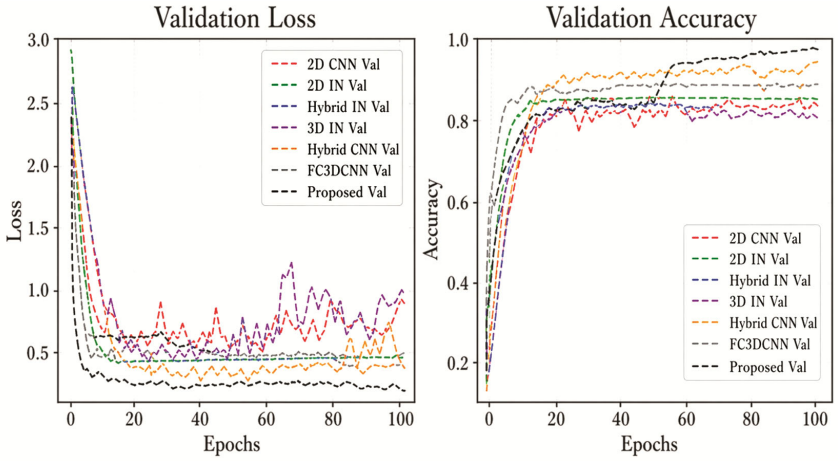
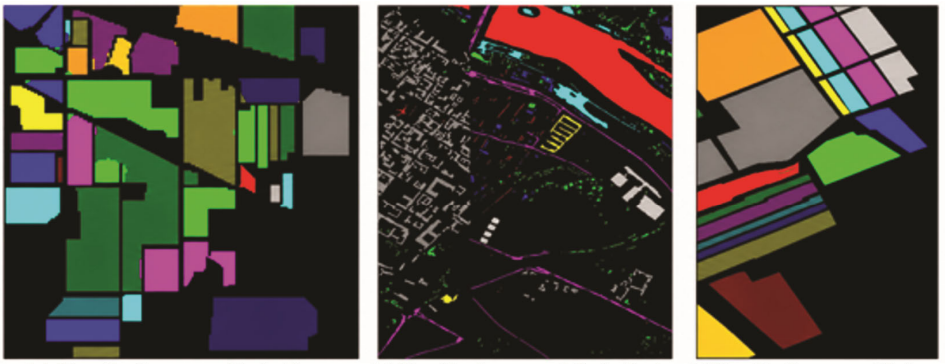


Fig. 3 — The Indian Pines dataset at 100 epochs validation loss and accuracy

Table 1 — Performance study of state-of-the-art (SOTA) models using class-wise F1-scores on the Indian Pines dataset

Classes	2D IN	2D CNN	3D IN	Hybrid CNN	Hybrid IN	FC3DCNN	Proposed Model
Alfalfa	66	48	84	67	80	72	100
Corn-notill	85	87	84	94	92	91	97
Corn-mintill	81	86	88	93	86	92	99
Corn	63	87	58	84	72	79	100
Grass-pasture	93	90	95	96	93	92	100
Grass-trees	97	97	97	98	98	96	98
Grass-mowed	57	81	87	87	96	98	96
Hay-windrowed	96	92	100	97	.95	99	100
Oats	80	67	36	87	80	88	100
Soybean-notill	89	92	86	98	94	95	99
Soybean-mintill	91	92	92	97	94	95	100
Soybean-clean	75	86	85	92	91	90	99
Wheat	87	91	98	95	99	94	100
Woods	98	94	97	98	96	95	100
Buildings	83	78	78	89	91	81	97
Stone-Steel	90	86	88	80	93	91	96
OA	91.16	89.23	88.84	96.26	91.87	93.28	98.98
AA	78.86	80.61	80.83	93.33	87.74	89.24	97.76
k	88.15	89.60	89.35	93.57	92.82	91.32	97.83



(a) Classification map of India Pines (OA:98.98%) (b) Classification map of Pavia Centre (OA:99.96%) (c)Classification map of Salinas (OA:99.92%)

Fig. 4 — Geographical depiction of classification results achieved by the proposed method

Table 2 — Performance study of state-of-the-art (SOTA) models using class-wise F1-scores on the Pavia Centre dataset

Classes	2D IN	2D CNN	3D IN	Hybrid CNN	Hybrid IN	FC3DCNN	Proposed Model
Water	100	100	100	100	100	100	100
Trees	99	99	98	99	99	100	100
Asphalt	98	98	94	93	98	98	99
Bricks	100	100	99	100	100	100	100
Bitumen	100	100	99	100	100	100	100
Tiles	100	100	99	100	100	100	100
Shadows	100	100	100	100	100	100	100
Meadows	100	98	100	97	100	99	100
Soil	100	100	97	98	98	100	100
OA	99.81	99.76	99.48	98.25	99.76	99.71	99.96
AA	98.52	98.18	99.59	99.08	98.47	98.55	99.88
k	98.01	99.69	99.28	99.17	99.66	99.32	99.93

model. The proposed model achieves OA at 98.98 %, which is significantly better than the others. As mentioned, the baseline models have OAs of 91.16 %, 89.23 %, 88.84 %, 96.26 %, 91.87 %, and 93.28 % for all other models, respectively. The superiority of the proposed model lies in its ability to perform joint modeling of spectral-spatial features and long-term spatial context dependency. The qualitative performance is shown in Fig. 4 (a).

Similarly, a performance evaluation of the proposed approach has been conducted for the Pavia Centre dataset. This is an urban scene and contains more LC classes compared with the Indian Pines dataset; Table 2 summarizes the classification performance of all the aforementioned models for each LC class. For the Water LC class, all models achieve an F1-score of 100 %, signifying its separability.

Regarding the Trees LC class, all models perform very well but the proposed model reaches the high value in terms of the F1-score, which indicates its superior capability in detecting spectral differences in vegetation LC classes[18]. On the other hand, some models, such as Hybrid IN and FC3DCNN, show relatively low performance in capturing the spectral properties of the complex urban environment. In the Soil class in the Pavia Centre dataset, all considered models, including SOTA and the Proposed Model, perform perfectly in terms of classification with an F1-score equal to 100 %. Such results unequivocally suggest that a high level of separability exists in the given case. The proposed model created the classification map shown in Fig. 4 (b).

When considering the execution of all models on the Salinas dataset, one can notice that almost all of them have similar accuracy[19,20]. The class-wise F1-scores for each model are presented in Table 3. Most models show high levels of separability with perfect F1-scores in most classes[21].

However, the proposed model performs even better with an OA of 99.92 %. This result confirms that the new model is capable of capturing more sophisticated spatial and spectral features of the land cover data[22]. All models showed excellent performance; however, the proposed one demonstrated consistent superiority in comparison with other models[23].

In the Salinas dataset performance analysis, each model achieved an OA score of approximately 99.2 %. Performance analysis shows a brief overview in Table 3, focusing on the F1 score for each class

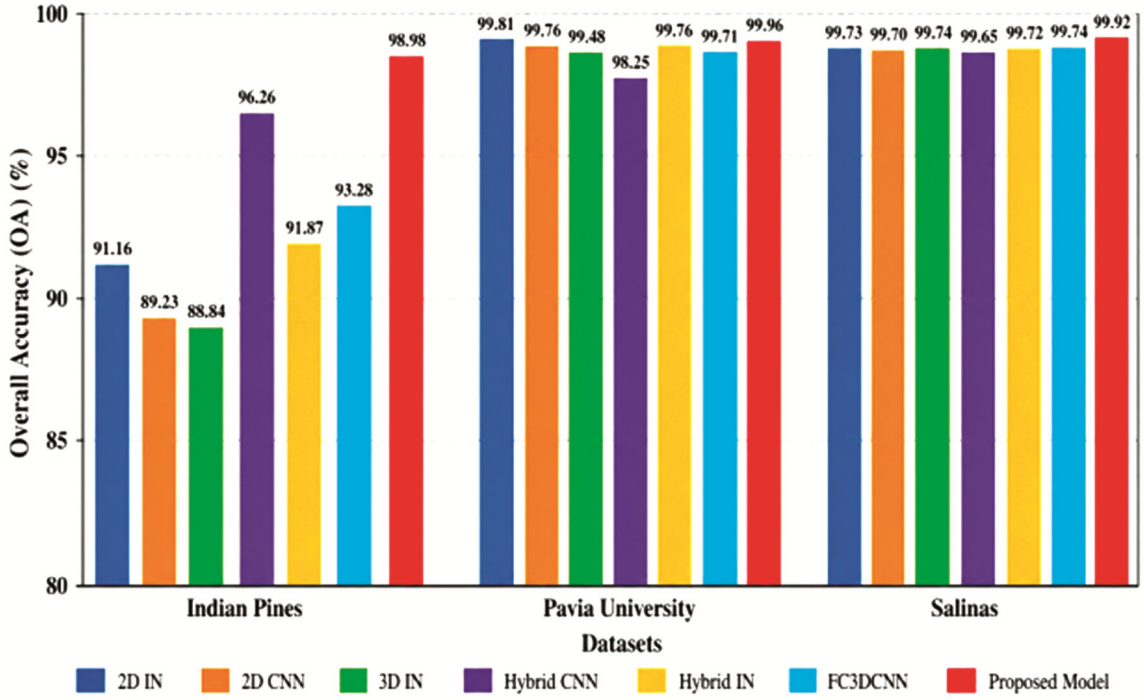


Fig. 5 — Comparison of overall accuracy (OA) for the proposed model hyperspectral image classification across all three datasets

achieved by the SOTA models for the Salinas dataset[24,25]. Interestingly, all the models show a perfect F1 score of 100 for almost all the LC classes. All the models show consistent accuracy, with the Proposed Model showing the highest accuracy with an impressive OA score of 99.92 %. for various LC classes included in the Salinas dataset, showing their high potential for accurately classifying complex features included in the agricultural dataset[26]. classification maps obtained from the Proposed Model, as illustrated in Fig. 4 (c). This graph shows a comparison of overall accuracy (OA) for the proposed hyperspectral image classification models against other traditional techniques in three standard datasets; namely Indian Pines, Pavia University, and Salinas. The proposed method is able to achieve higher OA than any other traditional method in all three datasets, with accuracies of 98.98 % in Indian Pines, 99.96 % in Pavia University, and 99.92 % in Salinas. In the Indian Pines dataset, the proposed technique proves to be much better compared to some of the traditional techniques like 2D IN (91.16 %), 2D CNN (89.23 %), and FC3DCNN (93.28 %) because of its better ability to learn spectral and spatial features.

All the techniques have achieved more than 98 % OA in both Pavia University and Salinas, but the proposed technique is still able to prove better than the rest. The OA comparison among all datasets is shown in Fig. 5.

Table 3 − Performance study of state-of-the-art (SOTA) models using class-wise F1-scores on the Salinas dataset

Classes	2D IN	2D CNN	3D IN	Hybrid CNN	Hybrid IN	FC3D CNN	Proposed Model
Weeds 1	100	100	100	100	100	100	100
Weeds 2	100	100	100	100	100	100	100
Fallow	100	100	100	100	100	100	100
Fallow Rough Plow	100	98	099	100	99	100	100
Fallow Smooth	100	99	100	100	100	100	100
Stubble	100	100	100	100	100	100	100
Celery	100	099	100	100	99	100	99
Grapes Untrained	100	100	100	99	100	99	100
Soil-Vinyard Develop	100	100	100	100	100	100	100
Corn Weeds	100	99	100	100	100	100	99
Lettuce 4wk	100	100	100	100	100	100	100
Lettuce 5wk	100	100	100	100	100	100	100
Lettuce 6wk	99	100	100	100	99	100	99

6 Conclusion

In light of the aforementioned constraints with the available models, our research fills the critical need for developing advanced techniques in land class classified in agricultural regions. The key aim is to

improve the classified accuracy and reliability by developing the Proposed Model with the incorporation of 3D-CNN GAT layers and VTs layers. The fusion of these two models helps to obtain an accurate understanding of the complex spatial relationships present in the HSI data. The classification accuracy achieved by the Proposed Model is significantly high compared to the other models for the Pavia Centre dataset, with an overall accuracy of 99.96 % compared to around 95 % for the other models. The practical implications of the Proposed Model are the improved precision and reliability in classifying LC in agricultural regions. The key advantage of the Proposed Model is the improved precision and reliability in classifying LC in agricultural regions. The key focus areas for future research include the exploration of the Proposed Model for other datasets, improved interpretability of the Proposed Model, and the extension of the Proposed Model for other remote sensing techniques. Moreover, the exploration of active learning as an important strategic direction for addressing the challenges related to the availability of labeled training data samples is also an important focus area for future research. The exploration of active learning techniques would focus on developing improvements in classification accuracy using sophisticated active learning algorithms.

Reference

1 Ji X, Cui Y, Wang H, Teng L, Wang L & Wang L, *IEEE Access*, 7 (2019) 146675.

2 Lou C, Al-qaness M A, AL-Alimi D, Dahou A, Abd Elaziz M, Abualigah L & Ewees A A, *Geo-spatial Inform Sci*, 28 (2) (2025) 345–386.

3 Shi S, Bi S, Gong W, Chen B, Tang X & Song S, *Remote Sens*, 13 (20) (2021) 4118.

4 Vali A, Comai S & Matteucci M, *Remote Sens*, 12 (15) (2020) 2495.

5 Babu R G, Geetha T S & Kumar K K, *Environm Monitor Assess*, 197 (11) (2025) 1275.

6 Liang M, Jiao L, Yang S, Liu F, Hou B & Chen H, *IEEE J Sel Top Appl Earth Observ Remote Sens*, 11 (8) (2018) 2911.

7 Wu K, Zhan Y, An Y & Li S, *Remote Sens*, 16 (13) (2024) 2328.

8 Cui X, Zheng K, Gao L, Zhang B, Yang D & Ren J, *Remote Sens*, 11 (19) (2019) 2220.

9 Gao H, Chen Z & Li C, *IEEE J Sel Top Appl Earth Observ Remote Sens*, 14 (2021) 5760.

10 Wang X & Fan Y, *IEEE J Sel Top Appl Earth Observ Remote Sens*, 15 (2022) 1617.

11 Shu Z, Zeng K, Zhou J, Wang Y, Tai M & Yu Z, *IEEE Trans Geosci Remote Sens*, 63 (2025) 5523316.

12 Yu H, Xu Z, Zheng K, Hong D, Yang H & Song M, *IEEE Trans Geosci Remote Sens*, 60 (2022) 1.

13 Li W, Yang Y, Zhang M, Mi P, Xiao Z & Xiang J, Deep spatial–spectral feature extraction network for hyperspectral image classification, In *Proceedings of the 43rd Chinese Control Conference*, *IEEE*, (2024) pp. 7955–7960.

14 Feng F, Zhang Y, Zhang J & Liu B, *Remote Sens*, 15 (2) (2023) 304.

15 Zhao X, Liang Y, Guo A J & Zhu F, *Remote Sens Lett*, 11 (4) (2020) 303.

16 Wu Y, Mu G, Qin C, Miao Q, Ma W & Zhang X, *Remote Sens*, 12 (1) (2020) 159.

17 Yang H, Yu H, Hong D, Xu Z, Wang Y & Song M, In *Proceedings of the 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing*, *IEEE*, (2022) pp. 1–4.

18 Guo T, Wang R, Luo F, Gong X, Zhang L & Gao X, *IEEE Trans Geosci Remote Sens*, 61 (2023) 1–13.

19 Liu, J., Wang, H., Liu, R., Wang, S., & Liu, Y. IEEE Trans Geosci Remote Sens, 62 (2024) 1.

20 Zhao C, Qin B, Feng S & Zhu W, *Remote Sens*, 14 (3) (2022) 681.

21 Li C, Wang Y, Fang Z & Li P, (2024) *IEEE Trans Geosci Remote Sens*, 62 (2024) 1.

22 Xie W, Zhang H & Zhang Y, *J Appl Remote Sens*, 19 (2) (2025) 024509.

23 Reddy C K K, Daduvy A, Mohana R M, Assiri B, Shuaib M, Alam S & Sheneamer M A, *IEEE Access*, 12 (2024) 125592.

24 Anuar M F A M S, Zaman F H K, Abdullah S A B C, Ng K M, Vijyakumar K N & Ng S K, *IEEE Open J Intell Transport Syst*, 7 (2026) 233.

25 Pennada S S P & Nayak S K, *Int Res J Multidisciplin Technov*, 7 (2) (2025) 74.

26 Bhukya S, Ganagoni V, Nangunoori S S & Enapothula S T, *Int Res J Multidiscipl Technov*, 7 (1) (2025).