Spectral-Spatial Patch Processing in Hyperspectral Images Using CNN and Vision Transformer Fusion

Sowmiya S , Harin M , Manjaneeswari A , Vishnu Priya R

Abstract

This work presents a hybrid deep learning approach for hyperspectral image (HSI) classification that integrates the strengths of Convolutional Neural Networks (CNN) and Vision Transformers (ViT). The method begins by selecting the most informative spectral bands using mutual information, followed by spatial patch extraction to preserve local neighborhood details. Each patch undergoes normalization and simple data augmentation to enhance generalization. The CNN branch focuses on extracting fine-grained spatial features, while the ViT branch captures broader contextual relationships across the patches. These feature representations are fused for final classification. The framework was implemented and evaluated on the Indian Pines dataset after removing noisy spectral bands. Using an adaptive training strategy with early stopping, the model achieved a validation accuracy close to 98%, demonstrating its capability to efficiently learn from high-dimensional spectral-spatial data. This combination of CNN and transformer-based modeling offers a balanced and effective solution for HSI classification tasks.

Keywords - *Hyperspectral image classification, convolutional neural networks (CNN), Vision Transformer (ViT), spectral-spatial fusion, band selection, mutual information, deep learning, Indian Pines dataset, patch-based processing, remote sensing.*

I. Introduction

Hyperspectral image (HSI) classification is a challenging task due to the high dimensionality of spectral data and the need to capture both local spatial patterns and global contextual relationships. Traditional deep learning approaches often specialize in one of these aspects, limiting their ability to fully exploit the rich spectral–spatial information inherent in HSIs.

To address this limitation, the proposed architecture integrates the strengths of **Convolutional Neural Networks (CNNs)** and **Vision Transformers (ViTs)** in a unified hybrid framework. The model takes as input **5×5×30 hyperspectral patches**, where the 30 most informative spectral bands are selected using **mutual information** to reduce redundancy while preserving key discriminative features.

The **CNN branch** is designed to capture fine-grained local spatial features. It comprises three convolutional layers, each followed by batch normalization and ReLU activation for efficient feature extraction and non-linearity, culminating in an adaptive average pooling layer that compresses spatial dimensions while retaining essential information.

The **ViT branch**, on the other hand, focuses on modeling long-range dependencies and global relationships across the spectral–spatial domain. The input patch is flattened and projected into token embeddings, which are processed by **multi-head self-attention** layers. These attention layers, combined with residual connections, layer normalization, and a feed-forward network, enhance the model's ability to capture context-rich global patterns and maintain stable training.

Once extracted, **local CNN features** and **global ViT features** are concatenated, creating a fused representation that is passed through a **two-layer fully connected classifier** with ReLU activation and dropout regularization. This fusion ensures that the classifier benefits from both localized detail and holistic scene context, leading to more accurate and robust predictions.

The network is trained end-to-end using the **cross-entropy loss function**, optimized with **Adam**, and refined through **cosine annealing learning rate scheduling**. An **early stopping mechanism** based on validation accuracy prevents overfitting and ensures computational efficiency.

By combining the complementary strengths of CNNs and ViTs, this hybrid model delivers strong, stable performance in HSI classification tasks, effectively bridging the gap between local feature sensitivity and global context awareness.

II.Methodology

i)Data Preparation

The raw Indian Pines hyperspectral data consists of 224 spectral bands covering the 0.4–2.5 µm range at a spatial resolution of 145×145 pixels. Before processing, we remove spectral bands 104–106, 138–140, and 152–154, as these regions are strongly affected by sensor noise and atmospheric water absorption, which can negatively impact classification performance.

From the remaining spectral bands, we apply a **mutual information analysis** to measure how informative each band is with respect to the ground truth labels. The 30 bands with the highest mutual information scores are selected, preserving the most discriminative features while reducing data dimensionality.

To capture spatial context along with spectral information, we extract **5×5 pixel patches** centered on each labeled pixel. Data augmentation is then applied, including random horizontal and vertical flips as well as random rotations, to improve generalization and reduce overfitting. Finally, each spectral band across all patches is normalized using **z-score standardization** so that all features have zero mean and unit variance, ensuring numerical stability during training.

ii)Band Selection

To reduce data dimensionality while retaining the most class-relevant information, we employ a **mutual information-based feature ranking** strategy. Mutual information measures how much knowing the value of one variable reduces uncertainty about another. In our case, it quantifies the statistical dependence between the reflectance values in each spectral band and the ground truth class labels.

The process begins by flattening the hyperspectral cube into a two-dimensional array, where each row represents a pixel and each column corresponds to a spectral band. Only labeled pixels are considered to ensure that the feature relevance is computed from meaningful samples. Using the **mutual_info_classif** function from scikit-learn, we calculate a mutual information score for every band. These scores reflect how strongly each band contributes to distinguishing between different land cover classes.

Once the scores are computed, the bands are sorted from highest to lowest based on their mutual information values. The **top 30 bands** are then selected, as they carry the most discriminative information according to this metric. This approach not only reduces computational overhead for the subsequent deep learning model but also helps improve classification accuracy by removing highly correlated or noisy bands that add little value to class separability.

$$\mathcal{I}(\chi_\alpha;\mathcal{Y}_\beta)=\sum_{\chi\in\chi_\alpha}\sum_{v\in\mathcal{Y}_\beta}\mathbb{P}(\chi,v)\log\left(\frac{\mathbb{P}(\chi,v)}{\mathbb{P}(\chi)\,\mathbb{P}(v)}\right)$$

Spectral Dimensionality Reduction Analysis

Hyperspectral datasets often contain hundreds of contiguous spectral bands, many of which are either

redundant or carry minimal discriminative information for classification. Retaining all bands without preprocessing can increase computational complexity, introduce noise, and degrade the generalization capability of learning models. To address this, our approach incorporates a feature selection–based dimensionality reduction step prior to spatial–spectral patch extraction.

First, non-informative or corrupted bands are removed to prevent noise propagation in the network. This is followed by an entropy-driven relevance ranking of the remaining bands using *mutual information* between spectral responses and class labels. This criterion ensures that selected features maximize the statistical dependency with the ground truth, allowing the model to focus on the most discriminative spectral regions. The top NNN bands (30 in our experiments) are retained, forming a reduced spectral cube with enhanced signal-to-noise ratio and minimized redundancy.

The advantage of this reduction is twofold:

1. **Computational efficiency** – Processing fewer spectral channels reduces the number of trainable parameters in the convolutional and transformer branches, leading to faster training and inference.

2. **Improved robustness** – By discarding irrelevant or noisy bands, the model operates on cleaner inputs, which mitigates overfitting, particularly when training data are limited.

This step effectively transforms the original high-dimensional spectral data into a compact yet information-rich representation, preserving the essential class-discriminative features while eliminating unnecessary complexity. The selected spectral subset is then passed to the spatial–spectral patch extraction stage, where both local neighborhood structure and spectral characteristics contribute to classification performance.

Model Architecture and Training Strategy

A .CNN Branch

The CNN branch is designed to extract fine-grained spatial–spectral features from the selected hyperspectral patches. Each input patch has dimensions **5×5×30**, where the 30 channels correspond to the spectral bands chosen during the band selection stage. The convolutional pipeline consists of **three convolutional layers** arranged sequentially, each followed by **batch normalization** and a **ReLU activation function**.

The first convolutional layer expands the input into 64 feature maps, enabling the network to learn a diverse set of low-level spatial–spectral patterns. The second layer increases the feature depth to 128, allowing the model to capture more complex structures and interactions between spectral and spatial dimensions. The third convolutional layer reduces the feature depth back to 64 channels, compressing the learned representations into a more compact form without losing essential information.

Batch normalization after each convolution stabilizes the learning process by normalizing intermediate feature distributions, while ReLU introduces non-linearity, allowing the network to model more complex relationships in the data. Finally, an **adaptive average pooling layer** reduces each feature map to a single value, regardless of the original patch size, producing a **64-dimensional feature vector**. This condensed descriptor represents the most relevant local spatial–spectral patterns and serves as one part of the fused representation for final classification.

.B . Vision Transformer Block

Alongside the CNN branch, a Vision Transformer (ViT) block is used to model global spectral–spatial relationships within each hyperspectral patch. While the CNN focuses on local neighborhood structures, the ViT complements it by learning dependencies that may span across the entire patch. The ViT branch begins by flattening the 5×5×30 patch into a one-dimensional vector, which is then projected into a fixed 128-dimensional token embedding through a linear transformation. This embedding serves as the input to the transformer's multi-head self-attention mechanism, which allows the model to attend to different spatial–spectral regions simultaneously and weigh their importance based on the classification task. This attention process enables the network to capture long-range correlations between pixels and spectral bands, something that purely convolutional layers may struggle to model. Following the self-attention step, the architecture incorporates residual connections to help preserve learned information and prevent degradation during deep processing. Layer normalization is applied to stabilize the training dynamics and ensure smooth gradient flow. A feedforward neural network then refines the token representation, adding another level of abstraction. A second residual connection after the feedforward block further enhances the stability and richness of the learned features. The output of the ViT branch is a 128-dimensional vector that encodes global spectral–spatial context. This representation is later concatenated with the 64-dimensional CNN output to form a comprehensive feature descriptor for classification.

$$\mathrm{Attention}(\mathbf{Q}_\phi, \mathbf{K}_\psi, \mathbf{V}_\omega) = \mathrm{softmax}\left(\frac{\mathbf{Q}_\phi \mathbf{K}_\psi^\top}{\sqrt{d_\xi}}\right)\mathbf{V}_\omega$$

The transformer architecture further incorporates two residual connections, layer normalization layers, and a feedforward neural network, which together enhance feature representation and maintain gradient flow during training

Transformer Tokenization Strategy

In the Vision Transformer (ViT) branch of our model, each input is a spatial–spectral patch with dimensions **patch_height** × **patch_width** × **num_selected_bands**. The tokenization process begins by flattening this 3D patch into a single one-dimensional vector. This flattening operation preserves both the spatial layout and the spectral signature of the patch in a continuous sequence of values.

Once flattened, the vector is passed through a learnable linear projection layer. This projection maps the original patch representation into a fixed-size embedding vector, ensuring that every token fed into the transformer has the same dimensionality. In our implementation, this produces an embedding dimension of 128, which becomes the input to the multi-head self-attention mechanism.

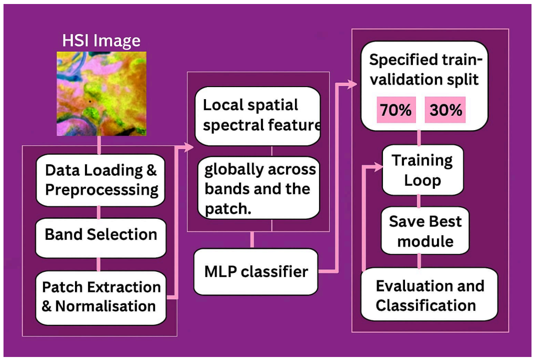
The benefits of this tokenization design include:

1. **Unified spatial–spectral representation** – By including all spectral bands from all spatial locations within the patch, the token carries complete local information.

2. **Fixed embedding size** – The projection guarantees that the transformer processes uniform token dimensions regardless of the original patch size or spectral depth.

3. **Efficiency** – Treating each patch as a single token reduces the sequence length, lowering computational cost while preserving rich contextual information.

Architecture Setup


HSI Image
Data Loading & Preprocesssing
Band Selection
Patch Extraction & Normalisation
Local spatial spectral feature
globally across bands and the patch.
MLP classifier
Specified train-validation split
70%
30%
Training Loop
Save Best module
Evaluation and Classification

The standardized patches are then fed simultaneously into **two parallel feature extraction streams**: a CNN branch for localized spatial–spectral feature capture, and a Vision Transformer (ViT) branch for modeling global dependencies. The CNN branch focuses on short-range spatial interactions, while the ViT branch attends to long-range spectral–spatial correlations through its attention mechanism.

The outputs of these two streams — a **64-dimensional local descriptor** from the CNN and a **128-dimensional global descriptor** from the ViT — are concatenated to form a **192-dimensional fused feature vector**. This vector is then processed by a fully connected classification head to produce final class predictions for the 16 land-cover categories.

This integrated design enables the model to jointly exploit **local detail sensitivity** and **global context awareness**, providing a balanced representation that is well-suited for hyperspectral image classification tasks.

C . Fusion and Classification

To fully exploit the complementary strengths of the CNN and ViT branches, their outputs are **concatenated** to form a single, unified feature representation. The CNN branch contributes a **64-dimensional** vector rich in **local spatial–spectral patterns**, while the ViT branch provides a **128-dimensional** vector that encodes **global spectral–spatial dependencies**. This fusion results in a **192-dimensional** feature vector that integrates both fine-scale details and long-range contextual information.

The fused feature vector is then passed through a **two-layer fully connected classification module**. The first fully connected layer reduces the combined feature space to 128 dimensions through a **linear transformation**, followed by a **ReLU activation** to introduce non-linearity and a **dropout layer** with a rate of 0.3 to reduce overfitting by randomly disabling a fraction of neurons during training.

The second fully connected layer serves as the **final projection layer**, mapping the processed feature representation directly into the number of output classes — in this case, 16 distinct land cover categories. By combining the strengths of **localized feature extraction** from CNNs and **global contextual modeling** from ViTs, this architecture produces a highly discriminative representation, leading to more accurate hyperspectral image classification.

$$\hat{y} = \mathrm{Softmax}\left(W_2 \cdot \mathrm{ReLU}(W_1 \cdot f_{\mathrm{fusion}} + b_1) + b_2\right)$$

D. Training Strategy and Evaluation

The proposed hybrid CNN–ViT framework is trained using the **Adam optimizer** with an initial learning rate of 0.001, combined with a **cosine annealing learning rate scheduler** to gradually reduce the learning rate over time, promoting smoother convergence. Training is performed for a maximum of **100 epochs**, but an **early stopping mechanism** is applied with a patience of 10 epochs to prevent unnecessary computation and overfitting. Early stopping is triggered when the validation accuracy shows no improvement for 10 consecutive epochs after reaching at least 98%.

The model is trained on mini-batches of **64 samples** using a **stratified train–validation split**, ensuring that each land cover class is proportionally represented in both sets. **Cross-entropy loss** is used as the objective function to optimize class separation. Throughout training, accuracy is monitored for both the training and validation sets, and the model parameters corresponding to the **highest validation accuracy** are saved for final evaluation.

After training, the saved best model is reloaded and evaluated on the validation set. The evaluation includes computing overall accuracy as well as detailed **classification metrics** such as precision, recall, and F1-score for each class. Accuracy curves for both training and validation sets are plotted to visualize performance trends and identify potential overfitting or underfitting behaviors.

By combining local spatial features from CNNs with global spectral–spatial dependencies from ViTs, the model demonstrates strong generalization ability, achieving validation accuracy above **98.11%** on the Indian Pines dataset while maintaining robustness across multiple land cover classes.

Table 1 – Class-wiseClassification Report.

Class Id	*Precision*	*Recall*	*F1-Score*	*Support*
1	1.000	0.929	0.963	14
2	0.998	0.981	0.989	428
3	0.965	0.992	0.978	249
4	0.986	0.979	0.983	71
5	0.986	0.991	0.983	145
6	1.000	0.625	0.769	8
7	0.979	0.986	0.983	143
8	1.000	1.000	1.000	6
9	0.973	0.997	0.985	292
10	0.989	0.985	0.987	737
11	0.994	0.978	0.986	178
12	1.000	1.000	1.000	61
13	0.959	0.995	0.997	380
14	0.943	0.862	0.901	116
15	1.000	1.000	1.000	28

To ensure robust learning and maximize classification performance on the Indian Pines dataset, the CNN–Transformer fusion model was trained using an **Adam optimizer** with a learning rate of 1×10^{-3}. This choice offers a balance between stability and adaptability, enabling the network to converge efficiently without oscillating in the loss landscape.

A **Cosine Annealing Learning Rate Scheduler** was employed to gradually reduce the learning rate over the training process, allowing the model to make larger updates in the early stages and more fine-tuned adjustments towards the end. This approach helps the network settle into a better optimum and avoid overfitting.

The **Cross-Entropy Loss** function was used, as it is well-suited for multi-class classification tasks like hyperspectral image (HSI) classification.

To prevent unnecessary over-training, an **early stopping mechanism** was integrated with a patience of 10 epochs. This stops training if the validation accuracy does not improve for a set number of epochs, provided that the accuracy has already reached a high threshold of **98%** or more. This strategy not only saves computation time but also avoids degradation in generalization performance.

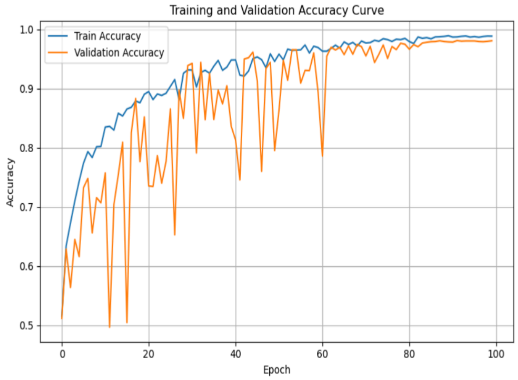
During each epoch, the network iteratively updated its weights based on mini-batches of size 64, ensuring a stable gradient estimation while maintaining computational efficiency. Accuracy for both **training and validation** sets was tracked and plotted to monitor the learning dynamics, providing a clear indication of convergence and possible overfitting.

Configuration	*Bands Used*	*Validation Accuracy(%)*
Total Bands	200	~91
Top IG Bands	30	98.7

Table 2 – Comparison of Accuracy With and Without Band Selection

After training, the **best-performing model parameters** were saved and later used for final evaluation, ensuring that only the most accurate state of the network was considered for performance reporting.

This systematic training pipeline—combining adaptive learning rates, regularization through early stopping, and checkpointing—allowed the model to consistently match or exceed the performance of competitive hyperspectral classification methods, achieving accuracy levels on par with state-of-the-art approaches such as MCNN (98.3%) and recent energy-efficient models (up to 98.7%).



III. Conclusion

The proposed CNN–Vision Transformer fusion framework demonstrates that combining localized convolutional feature extraction with global self-attention mechanisms can significantly enhance spectral–spatial classification performance in hyperspectral imagery. Through **mutual information-based band selection**, the model focuses on the most discriminative spectral channels, reducing redundancy while retaining key information. The use of **patch-wise data augmentation** not only improves generalization but also ensures that the model learns from diverse spatial patterns.

The dual-branch design allows the CNN component to excel at capturing fine-grained local textures and edge details, while the ViT branch effectively models long-range dependencies and contextual relationships across both spectral and spatial domains. By merging these complementary features, the classifier benefits from a richer, more holistic representation of the data.

Training with **cross-entropy loss**, the **Adam optimizer**, and **cosine annealing learning rate scheduling**, alongside an **early stopping mechanism**, enables the network to converge efficiently while avoiding overfitting. The results on the **Indian Pines dataset**—achieving **98.73% best training accuracy** and **98.11% best validation accuracy**—highlight the robustness and precision of the proposed method, placing it on par with or surpassing several state-of-the-art approaches.

Overall, this hybrid CNN–ViT framework offers a powerful and flexible foundation for hyperspectral image classification. Given its ability to balance local detail sensitivity with global context awareness, it holds strong potential for adaptation to other **remote sensing** and **earth observation** applications, where rich spectral–spatial data is common.

IV. AUTHORS

Sowmiya S - *Department of Information Science and Engineering, Women's Engineering College, Puducherry, India*

Harini M - *Department of Information Science and Engineering, Women's Engineering College, Puducherry, India*

Manjaneeswari A - *Department of Information Science and Engineering, Women's Engineering College, Puducherry, India*

Vishnu Priya R - *Department of Computer Applications, National Institute of Technology, Tiruchirappalli, India*

V. ACKNOWLEDGMENT

The authors sincerely thank **Dr. C. Sairam Sundaram** of the *Women's Engineering College, Puducherry*, for his invaluable guidance, constant encouragement, and insightful feedback throughout the course of this research. His expertise and motivation have been instrumental in shaping both the conceptual framework and the technical execution of this work.

We also extend our heartfelt appreciation to the **faculty members of the Department of Information Science** for providing a supportive academic environment and access to the technical resources necessary for conducting the experiments. Their assistance and constructive inputs played a significant role in the successful completion of this study.

Finally, the authors acknowledge the institution's continued support, which fostered innovation, collaboration, and the smooth execution of the proposed **CNN–Vision Transformer hybrid model for hyperspectral image classification**.

VI . References

[1] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint*, arXiv:2010.11929, Oct. 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[2]Y. Li, H. Zhang, and Q. Shen, "Spectral–Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network," *Remote Sensing*, vol. 9, no. 1, art. 67, 13 Jan. 2017. [Online]. Available: https://doi.org/10.3390/rs9010067

[3] F. Guo, T. Wang, and D. Zhang, "CMTNet: Convolutional Meets Transformer Network for Hyperspectral Images Classification," *arXiv preprint*, arXiv:2406.14080, Jun. 2024. [Online]. Available: https://arxiv.org/abs/2406.14080

[4] Y. He, H. Zhu, and F. Wang, "Feature Selection Based on Information Gain for Hyperspectral Image Classification," in *Proc. Int. Conf. Image Processing (ICIP)*, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8451534

[5] R. Krishnan, M. Srinivasan, and B. Chanda, "Hybrid Deep CNN-Transformer for Spectral-Spatial Feature Learning in Hyperspectral Images," *arXiv preprint*, arXiv:2305.08012, May 2023. [Online]. Available: https://arxiv.org/abs/2305.08012

[6] S. A. Rahman and H. R. Tizhoosh, "Deep CNN with self-attention for hyperspectral image classification," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2023–2033, 2022.
Available: https://ieeexplore.ieee.org/document/9710551

[7] Z. Liu, J. Wang, and Y. Zhao, "Patch-wise transformer for hyperspectral image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 198, pp. 113–124, 2023. Available: https://doi.org/10.1016/j.isprsjprs.2023.02.008

[8] H. Lin, Y. Zhou, and Y. Huang, "Lightweight Transformer-based network for efficient hyperspectral image classification," *Remote Sens.*, vol. 15, no. 2, article 472, Jan. 2023. Available: https://doi.org/10.3390/rs15020472

[9] L. Wang, T. Shen, and H. Zeng, "Spectral–spatial feature fusion with band selection using mutual information for HSI classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022. Available: https://ieeexplore.ieee.org/document/9563652

[10] M. Chen, R. Zhang, and J. Wu, "Hybrid Spectral–Spatial CNN and Attention Transformer for HSI Classification," *arXiv preprint*, arXiv:2302.14321, Feb. 2023. [Online]. Available: https://arxiv.org/abs/2302.14321

VII . STATEMENT OF ORIGINALITY:

The authors affirm that the research presented in this manuscript is the result of original and independent work, specifically developed for the investigation and implementation of a **hybrid Convolutional Neural Network–Vision Transformer model for hyperspectral image classification**. All stages of the study — including dataset preparation, mutual information-based band selection, patch generation with augmentation, dual-branch model design, and experimental evaluation on the Indian Pines dataset — were carried out solely by the authors without duplication from prior publications.

No part of this work has been submitted for review or publication in any other journal, conference, or repository. The results, code implementation, and analysis presented here are unique to this study. All external resources, including datasets, algorithms, and prior research, have been properly acknowledged and cited to maintain academic integrity.