Spectral–Spatial Hyperspectral Image Classification Using Machine Learning

Indian Pines Dataset

Saichandana Thumma
Indian Institute of Technology, Jodhpur
saichandanathumma029@gmail.com

December 17, 2025

Abstract

Hyperspectral imaging provides rich spectral information across hundreds of wavelength bands, enabling precise material and land-cover identification beyond conventional RGB imaging. However, the high dimensionality of hyperspectral data and strong class imbalance make accurate classification challenging.

In this work, we study hyperspectral image classification on the Indian Pines dataset and propose an improved spectral–spatial machine learning pipeline. A baseline spectral classifier using Principal Component Analysis (PCA) and Support Vector Machines (SVM) is first established. To enhance performance, spatial context is incorporated by augmenting each pixel's spectral signature with the mean spectrum of its local neighborhood.

Experimental results demonstrate that spectral–spatial feature fusion significantly improves classification performance, increasing balanced accuracy from 84.2% to 89.3% and macro F1-score from 0.78 to 0.83. The proposed approach produces coherent classification maps with reduced noise and sharper class boundaries, highlighting the importance of spatial information in hyperspectral image analysis.

Contents

1 Introduction

Hyperspectral imaging (HSI) has emerged as a powerful sensing technology due to its ability to capture detailed spectral information across hundreds of contiguous wavelength bands. Unlike conventional RGB imaging, which is limited to three broad color channels, hyperspectral images provide a full reflectance spectrum for every pixel, enabling fine-grained discrimination of materials and land-cover types. This capability has made hyperspectral imaging particularly valuable in applications such as agriculture, environmental monitoring, mineral exploration, medical diagnostics, and remote sensing.

Despite its advantages, hyperspectral image classification presents several significant challenges. The extremely high dimensionality of hyperspectral data often leads to increased computational complexity and the curse of dimensionality, making learning difficult for traditional machine learning algorithms. Additionally, real-world hyperspectral datasets frequently exhibit strong class imbalance, where certain land-cover classes contain only a small number of labeled samples. Spectral similarity between different classes, such as crops with similar vegetation characteristics, further complicates accurate classification.

To address these challenges, dimensionality reduction and robust classification techniques are commonly employed. Principal Component Analysis (PCA) is widely used to reduce spectral dimensionality while preserving most of the informative variance in the data. Support Vector Machines (SVMs) have also proven effective for hyperspectral classification due to their strong generalization ability in high-dimensional feature spaces, particularly when training data is limited.

However, purely spectral approaches treat each pixel independently and ignore the spatial structure inherent in hyperspectral images. In practice, neighboring pixels often belong to the same class and share similar spectral characteristics. Incorporating spatial context can therefore significantly improve classification robustness by reducing noise and enhancing class separability.

In this work, we investigate hyperspectral image classification on the Indian Pines dataset and propose an improved spectral–spatial machine learning pipeline. A baseline spectral classifier using PCA and SVM is first established. Spatial information is then integrated by augmenting each pixel's spectral signature with the mean spectrum of its local neighborhood. Experimental results demonstrate that this spectral–spatial fusion leads to substantial performance improvements, producing more coherent classification maps and improved accuracy across both majority and minority classes.

2 Dataset Description

Dataset link: `https://www.kaggle.com/datasets/abhijeetgo/indian-pines-hyperspectral-dataset`
The experiments in this study are conducted on the widely used Indian Pines hyperspectral dataset, which is a benchmark dataset for evaluating hyperspectral image classification algorithms. The dataset was captured by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over agricultural areas in northwestern Indiana, USA.

The hyperspectral image consists of spatial dimensions of 145×145 pixels with 200 spectral bands after the removal of noisy and water-absorption bands. Each pixel is represented by a 200-dimensional spectral vector, corresponding to reflectance values measured across contiguous wavelength bands in the visible and infrared regions of the electromagnetic spectrum.

A ground truth label map is provided for supervised learning, where each labeled pixel is assigned to one of 16 land-cover classes, primarily consisting of different crop types, vegetation, and man-made structures. The dataset exhibits strong class imbalance, with some classes such as Soybean-mintill and Corn-notill containing thousands of samples, while others such as Oats and Grass-pasture-mowed contain fewer than 30 labeled pixels. This imbalance poses a significant challenge for classification algorithms and makes the dataset particularly suitable for

evaluating robustness to minority classes.

Pixels labeled as background or unknown are excluded from training and evaluation, resulting in a total of 10,249 labeled samples used in the experiments. The diversity of land-cover types, high spectral similarity between certain crop classes, and severe imbalance together make the Indian Pines dataset a challenging and representative benchmark for hyperspectral image analysis.

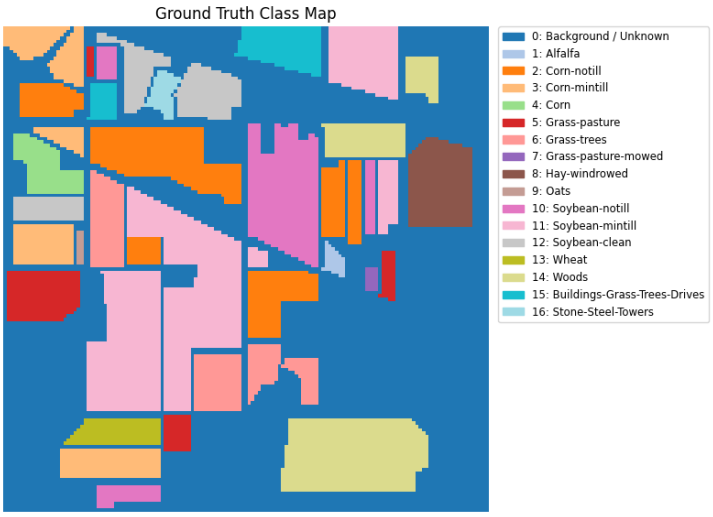

Figure 1: Ground truth class map of the Indian Pines hyperspectral dataset.

3 Methodology

This section presents the proposed hyperspectral image classification pipeline, including data representation, preprocessing, baseline spectral modeling, and spectral–spatial feature fusion. The overall objective is to perform accurate pixel-wise classification by exploiting both spectral signatures and local spatial context.

3.1 Hyperspectral Image Representation

A hyperspectral image is represented as a three-dimensional data cube:

$$\mathbf{X} \in \mathbb{R}^{H \times W \times B},$$

where H and W denote the spatial dimensions (height and width), and B denotes the number of spectral bands. Each pixel located at spatial position (i, j) is associated with a spectral vector:

$$\mathbf{x}_{i,j} = [x_{i,j}^{(1)}, x_{i,j}^{(2)}, \ldots, x_{i,j}^{(B)}]^\top,$$

where $x_{i,j}^{(b)}$ represents the reflectance value at the b-th wavelength.

In the Indian Pines dataset, the hyperspectral cube has dimensions $145 \times 145 \times 200$, meaning that each pixel is characterized by a 200-dimensional spectral signature. Pixel-wise hyperspectral classification aims to assign a land-cover class label $y_{i,j}$ to each pixel based on its spectral and spatial characteristics.

3.2 Feature Matrix Construction

For supervised learning, the hyperspectral cube is reshaped into a two-dimensional feature matrix:

$$\mathbf{X}_{\text{spec}} \in \mathbb{R}^{N \times B},$$

where $N = H \cdot W$ is the total number of pixels, and each row corresponds to the spectral vector of a single pixel. The corresponding ground truth labels are represented as:

$$\mathbf{y} \in \{0, 1, \ldots, C\}^N,$$

where C is the number of land-cover classes and label 0 denotes background pixels.

Pixels labeled as background are excluded from training and evaluation, resulting in a reduced dataset containing only meaningful class samples.

3.3 Data Preprocessing and Normalization

Due to large variations in reflectance values across spectral bands, feature standardization is applied to ensure numerical stability and balanced feature contribution. Each spectral band is standardized using:

$$x' = \frac{x - \mu}{\sigma},$$

where μ and σ denote the mean and standard deviation computed from the training data.

To address the strong class imbalance present in the dataset, a stratified train-test split is employed, ensuring that class proportions are preserved in both training and testing sets.

3.4 Baseline Spectral Classification

As a baseline model, spectral-only classification is performed using Principal Component Analysis (PCA) followed by a Support Vector Machine (SVM) classifier.

PCA projects the original spectral features onto a lower-dimensional subspace:

$$\mathbf{Z} = \mathbf{X}_{\text{spec}}\mathbf{W},$$

where $\mathbf{W}$ contains the eigenvectors corresponding to the largest eigenvalues of the data covariance matrix. In this work, the first 30 principal components are retained, preserving approximately 98.6% of the total spectral variance.

An SVM with a radial basis function (RBF) kernel is trained on the reduced feature space. To mitigate the effects of class imbalance, class-weighted learning is employed, assigning higher penalties to misclassification of minority classes. Model performance is evaluated using balanced accuracy and macro-averaged F1-score.

3.5 Spectral–Spatial Feature Fusion

While spectral-only approaches treat each pixel independently, hyperspectral images exhibit strong spatial continuity, as neighboring pixels often belong to the same land-cover class. To exploit this spatial structure, spectral–spatial feature fusion is introduced.

For each pixel (i, j), a local neighborhood $\mathcal{N}_{i,j}$ of size 3×3 is defined. The neighborhood mean spectral vector is computed as:

$$\bar{\mathbf{x}}_{i,j} = \frac{1}{|\mathcal{N}_{i,j}|} \sum_{(p,q) \in \mathcal{N}_{i,j}} \mathbf{x}_{p,q}.$$

The final fused feature vector is obtained by concatenating the original spectral vector with its neighborhood mean:

$$\mathbf{x}_{i,j}^{\text{fused}} = [\mathbf{x}_{i,j}; \bar{\mathbf{x}}_{i,j}] \in \mathbb{R}^{2B}.$$

PCA is reapplied to the fused feature space to reduce dimensionality, followed by SVM classification. This spectral–spatial representation improves robustness to noise, enhances class separability, and yields more coherent classification results.

4 Experiments and Results

This section presents the experimental setup, evaluation metrics, and classification results obtained using the proposed hyperspectral image classification pipeline. Baseline spectral models are first evaluated, followed by an analysis of the improved spectral–spatial approach.

4.1 Experimental Setup

All experiments are conducted on the Indian Pines hyperspectral dataset after removing background pixels. The dataset is split into training and testing sets using a stratified sampling strategy, with 80% of labeled pixels used for training and 20% for testing. Feature standardization is applied prior to dimensionality reduction and classification.

Principal Component Analysis (PCA) is used to reduce spectral dimensionality, retaining 30 components that preserve approximately 98.6% of the total variance. Classification is performed using Support Vector Machines (SVM) with a radial basis function kernel and class-weighted learning to address class imbalance. A K-Nearest Neighbors (KNN) classifier is also evaluated as a comparative baseline.

Model performance is evaluated using balanced accuracy and macro-averaged F1-score, which provide robust assessment under class imbalance. Per-class precision, recall, and F1-scores are additionally analyzed.

4.2 Spectral Signature Analysis

To illustrate the discriminative power of hyperspectral data, spectral signatures from representative pixels belonging to different land-cover classes are visualized. Figure 2 shows reflectance curves across 200 spectral bands for selected classes.

Although certain crop types appear visually similar in RGB imagery, their hyperspectral signatures exhibit distinct patterns across specific wavelength regions. Differences in reflectance within the visible, red-edge, and near-infrared regions enable effective separation of spectrally similar classes, highlighting the advantage of hyperspectral imaging over conventional RGB-based approaches.

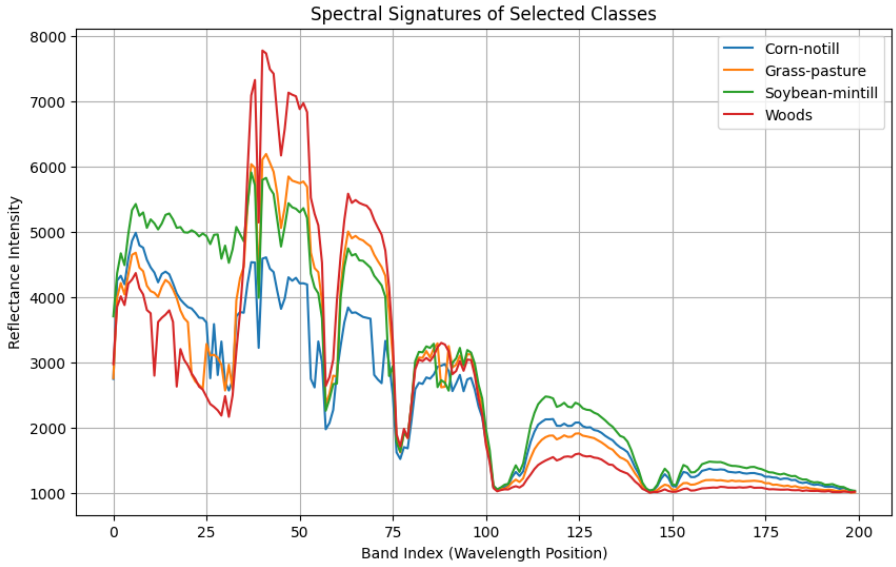

Figure 2: Spectral signatures of representative pixels from different land-cover classes across 200 hyperspectral bands.

4.3 Baseline Classification Results

A baseline spectral-only classification is performed using PCA followed by SVM. The baseline model achieves a balanced accuracy of 84.2% and a macro F1-score of 0.776. Strong performance is observed for classes with distinctive spectral characteristics such as Woods, Wheat, and Hay-Windrowed. However, classes with high spectral similarity, including different corn and soybean variants, exhibit comparatively lower F1-scores.

A KNN classifier is also evaluated as a simple baseline. While KNN performs reasonably well for classes with clear spectral separation, its overall performance is inferior to the SVM-based approach, particularly for minority and overlapping classes. This behavior is attributed to KNN's sensitivity to noise and class imbalance.

4.4 Spectral–Spatial Classification Results

To improve classification robustness, spatial context is incorporated by augmenting each pixel's spectral vector with the mean spectrum of its local 3×3 neighborhood. The resulting spectral–spatial features are classified using PCA followed by SVM.

This approach significantly improves performance, increasing balanced accuracy from 84.2% to 89.3% and macro F1-score from 0.776 to 0.832. Figure 3 summarizes the performance comparison between baseline and spectral–spatial models.

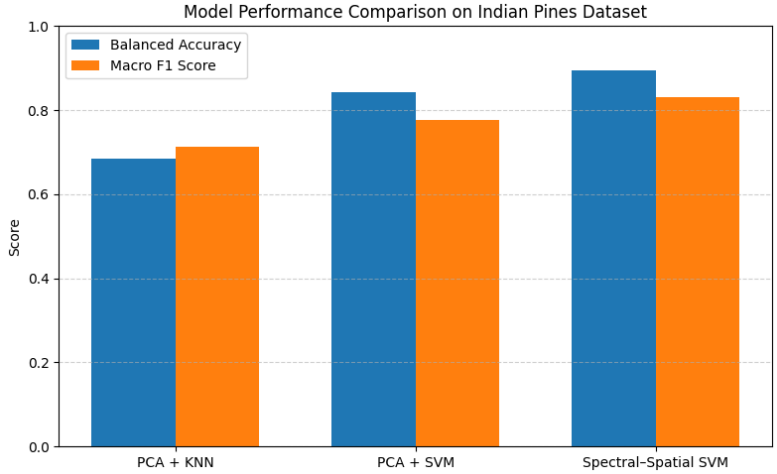

Figure 3: Comparison of balanced accuracy and macro F1-score between baseline spectral and spectral–spatial classification models.

Notable improvements are observed for spectrally similar crop classes such as Corn-notill, Corn-mintill, and Soybean variants. Incorporating spatial context reduces pixel-level noise and enhances class separability, particularly for minority classes.

Model	**Balanced Accuracy (%)**	**Macro F1-score**
KNN (Spectral Only)	68.6	0.714
PCA + SVM (Spectral Only)	84.2	0.776
Spectral–Spatial PCA + SVM	**89.3**	**0.832**

Table 1: Performance comparison of different classification approaches on the Indian Pines dataset

4.5 Classification Map Visualization

Figure 4 presents the full predicted classification map generated using the spectral–spatial SVM model. The predicted map exhibits coherent field-level structures with smooth regions and sharp boundaries between land-cover classes, indicating effective exploitation of spatial context.

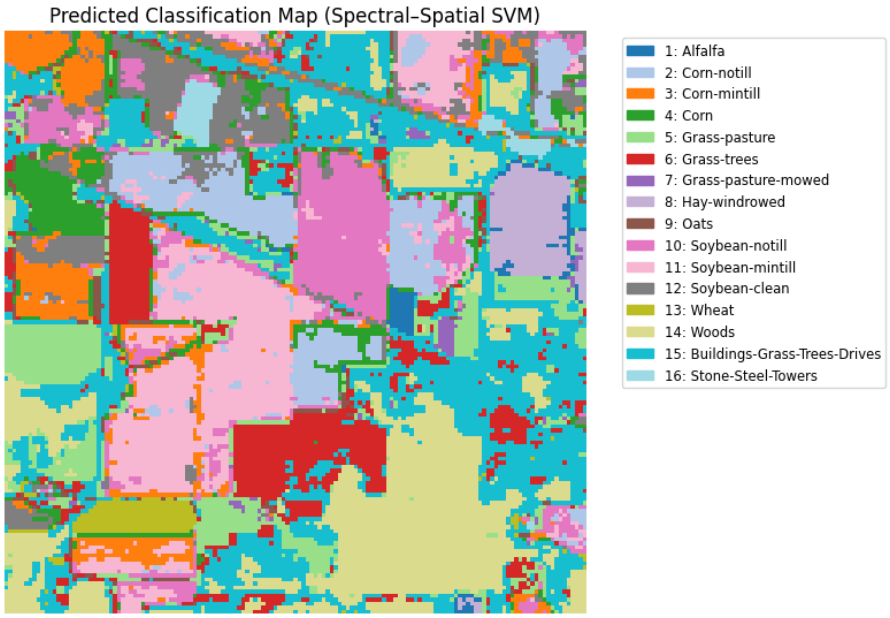

Figure 4: Predicted classification map produced by the spectral–spatial SVM model.

A visual comparison between the ground truth and predicted maps is shown in Figure 5. The predicted labels closely follow the spatial distribution of the ground truth, with most errors occurring within regions of spectrally similar crop types. This behavior is consistent with expectations for real-world hyperspectral classification tasks.

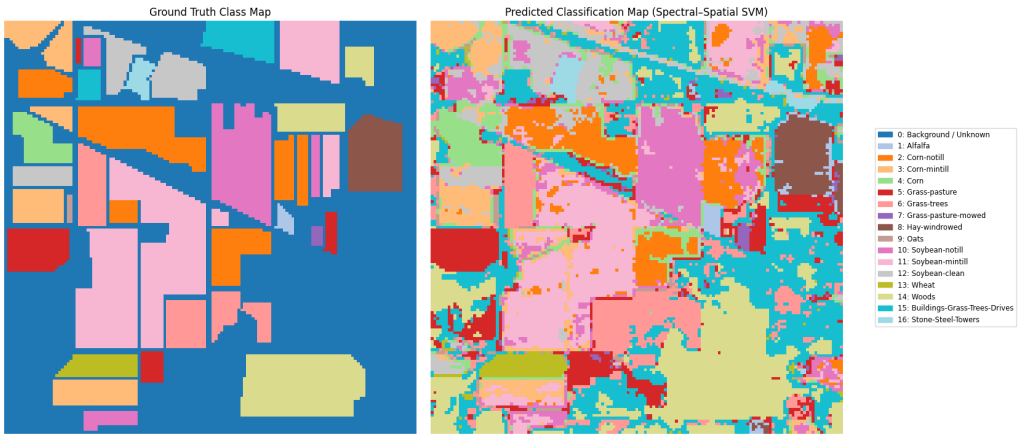

Figure 5: Side-by-side comparison of ground truth and predicted classification maps.

5 Discussion

The experimental results demonstrate that hyperspectral image classification benefits significantly from the integration of spatial context alongside spectral information. While the baseline spectral-only PCA–SVM model already achieves strong performance, its limitations become ev-

ident for classes with high spectral similarity and for minority classes with limited training samples.

The spectral signature analysis confirms that many agricultural classes in the Indian Pines dataset exhibit overlapping reflectance patterns in certain wavelength regions, particularly among different corn and soybean variants. As a result, purely spectral classifiers may struggle to establish clear decision boundaries when relying on pixel-wise information alone. This behavior is reflected in the baseline results, where spectrally similar crop types show comparatively lower F1-scores.

Incorporating spatial information through spectral–spatial feature fusion substantially improves classification robustness. By augmenting each pixel's spectral vector with the mean spectrum of its local neighborhood, the classifier is able to exploit spatial continuity present in agricultural fields. This reduces the impact of sensor noise and isolated pixel-level variations, leading to smoother classification maps and improved separability between similar classes. The observed increase in balanced accuracy and macro F1-score highlights the effectiveness of this approach, particularly in the presence of strong class imbalance.

The comparison between SVM and KNN classifiers further emphasizes the importance of model selection for hyperspectral data. While KNN performs reasonably well for classes with clear spectral separation, it is more sensitive to noise and class imbalance, resulting in lower overall performance. In contrast, the SVM classifier demonstrates greater robustness in high-dimensional feature spaces and provides more consistent performance across both majority and minority classes.

Despite the improvements achieved through spectral–spatial fusion, certain challenges remain. Misclassifications are still observed primarily among crop types with extremely similar spectral and spatial characteristics. Additionally, classes with very few labeled samples introduce uncertainty in per-class performance estimates, as small test set sizes can lead to higher variance in evaluation metrics. These limitations are inherent to the dataset and highlight the difficulty of hyperspectral classification in real-world scenarios.

Overall, the results indicate that combining spectral and spatial information is a practical and effective strategy for improving hyperspectral image classification, particularly for applications involving noisy data and imbalanced class distributions.

6 Conclusion and Future Work

In this work, a spectral–spatial machine learning pipeline for hyperspectral image classification was developed and evaluated using the Indian Pines dataset. A baseline spectral classifier based on PCA and SVM demonstrated strong performance, confirming the effectiveness of dimensionality reduction and margin-based learning for high-dimensional hyperspectral data. To address the limitations of spectral-only classification, spatial context was incorporated through local neighborhood feature fusion.

The proposed spectral–spatial approach significantly improved classification performance, increasing balanced accuracy from 84.2% to 89.3% and macro F1-score from 0.776 to 0.832. Visual inspection of the classification maps further confirmed that spatial feature integration produces smoother, more coherent predictions with reduced noise and improved separation of spectrally similar crop types. These results highlight the importance of leveraging both spectral signatures and spatial structure for robust hyperspectral analysis.

Future work may explore more advanced spatial representations, such as multi-scale neighborhoods, texture-based features, or morphological profiles, to further enhance class discrimination. Additionally, extending the framework to deep learning architectures, including convolutional or hybrid spectral–spatial neural networks, could enable end-to-end feature learning from raw hyperspectral data. Applying the proposed approach to medical hyperspectral imaging and satellite-based Earth observation data represents another promising direction, aligning

with real-world diagnostic and remote sensing applications.

References

[1] Indian Pines Hyperspectral Dataset. https://www.kaggle.com/datasets/abhijeetgo/indian-pines-hyperspectral-dataset

[2] Hyperspectral Remote Sensing Scenes. https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes