Robust Hyperspectral Image Classification: A Transformer-Based Approach with Enhanced Adversarial Defense

Joshua Orndorff and Josue Syvelsaint
Department of Computer Science and Engineering
Fairfield University
Emails: joshua.orndorff@student.fairfield.edu, josue.syvelsaint@student.fairfield.edu

***Abstract*—Hyperspectral image classification has become crucial for remote sensing applications, but existing systems are vulnerable to adversarial attacks that undermine their reliability. This research introduces a novel approach that integrates transformer-based architectures with multi-scale feature extraction and a custom loss function incorporating adversarial training. The proposed method leverages spectral and spatial attention mechanisms to capture complex relationships within hyperspectral data while enhancing model robustness against adversarial perturbations. Experimental results show that our model achieves high classification accuracy on clean data and demonstrates improved performance in adversarial settings, effectively minimizing errors caused by attacks.**

I. Introduction

Hyperspectral image (HSI) classification has seen a breakthrough in recent years driven by the advancement of deep learning frameworks [1], this progress has been marked by the high precision of prediction using minimal training samples. Traditional methods often struggled with the large volume of spectral bands inherent in HSI [2], but deep learning, particularly convolutional neural networks (CNNs) [3] and transformer-based architectures, has demonstrated the ability to learn complex, high-level representations from this data.

Hyperspectral imaging involves the collection and processing of information across a wide range of the electromagnetic spectrum, enabling the identification and classification of materials based on their unique spectral signatures. This technology is widely used in applications such as environmental monitoring, agriculture, mineral exploration, and defense due to its ability to provide detailed and accurate analysis of surface composition and features [4].

The SpectralFormer [5] model represents a significant advancement in HSI analysis, leveraging the power of transformer-based architectures to effectively process high-dimensional spectral data. Unlike traditional CNNs, SpectralFormer uses self-attention mechanisms to capture both local and global relationships within the data, enabling more nuanced feature extraction. Its architecture is specifically designed to accommodate the unique characteristics of HSI, their extensive spectral bands, and interband correlations. By integrating spectral and spatial information, the model achieves state-of-the-art performance in classification tasks, making it a robust choice for applications requiring precise and reliable analysis.

The discovery of adversarial vulnerabilities in deep neural networks (DNNs) marked a significant turning point in our understanding of artificial intelligence systems [6]. Research demonstrated that these sophisticated models, despite their impressive performance, harbor a critical weakness: Their predictions can be manipulated by introducing carefully crafted, minor perturbations to the input data [7]. These perturbations, while imperceptible to human observers, can cause state-of-the-art models to make dramatically incorrect predictions with high confidence.

This vulnerability has particular implications for HSI classification systems, where the stakes of misclassification can be especially high. In HSI applications, these adversarial manipulations can be even more subtle than in traditional image classification tasks, as they can be distributed across multiple spectral bands that lie outside the human visual perception [8]. The complexity of hyperspectral data, with hundreds of spectral bands, provides numerous opportunities for adversarial manipulation while making detection of such alterations extremely challenging.

Given the critical nature of HSI classification in remote sensing applications, the potential vulnerability to adversarial attacks represents a significant concern that demands attention. Our research is motivated by the growing deployment of these classification systems in security-sensitive domains, where reliability and robustness are paramount. Whether monitoring agricultural conditions, conducting environmental assessments or supporting defense applications, the integrity of these systems must be maintained even in the presence of potential adversarial manipulation.

The intersection of transformer architectures and adversarial defense mechanisms presents a compelling opportunity to address these challenges. Although transformers have shown remarkable success in processing sequential data and capturing long-range dependencies, their application to robust HSI classification remains an area ripe for investigation. By combining the strengths of transformer-based architectures with targeted defense strategies, our goal is to develop a more resilient system that maintains high performance under adverse conditions.

II. Dataset

We evaluated our method on three standard hyperspectral datasets[1]: Indian Pines, University of Pavia, and Salinas Scene. Indian Pines and Salinas Scene contain agricultural scenes with crop-type labels, while University of Pavia captures an urban area with labels ranging from buildings to bare soil. The Indian Pines dataset has a spatial dimension of 145×145 with 200 spectral bands. University of Pavia consists of 1096×715 pixels with 102 spectral bands. Salinas Scene has dimensions of 512×217 with 204 spectral bands.

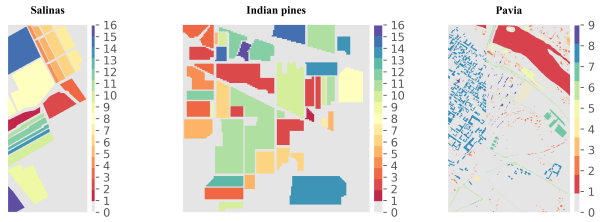

Fig. 1. Ground Truth Labels for each of the three images

Each dataset underwent standardization across spectral bands by flattening the spatial dimensions, applying standard scaling, and reshaping back to the original dimensions. We then constructed spatial-spectral patches of size $R^{7\times7\times B}$, where B denotes the number of spectral bands. Each patch was labeled according to its center pixel, excluding background classes and padding edge pixels where necessary. Patches received random flips and rotations to limit the potential for data leakage.

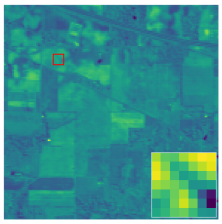

Fig. 2. Indian Pines dataset with an example of an extracted patch

The datasets were divided using a stratified sampling approach to maintain class distribution across the splits. For each class in all datasets, the data was partitioned into three subsets: training, validation, and testing sets, following a ratio of 1:1:3. This stratified approach ensures that the relative proportion of samples for each class is preserved across all sets, which is crucial for maintaining representative class distributions during model training and evaluation. As an example, Table I shows the distribution for the University of Pavia dataset, where the training and validation sets contain equal amounts of samples, while the test set contains approximately three times more samples for each class. This same sampling strategy was consistently applied across all datasets used in this study.

[1]https://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

TABLE I
EXAMPLE OF STRATIFIED SAMPLING DISTRIBUTION USING THE UNIVERSITY OF PAVIA DATASET

Label	Train	Validate	Test
Asphalt	13,194	13,194	39,583
Meadows	1,519	1,519	4,560
Gravel	618	618	1,854
Trees	537	537	1,611
Painted metal sheets	1,316	1,316	3,952
Bare Soil	1,849	1,849	5,550
Bitumen	1,457	1,457	4,373
Self-blocking bricks	8,565	8,565	25,696
Shadows	572	577	1,719

III. Methods

A. Model

To build a robust Hyperspectral Imaging (HSI) classification model, the SpectralFormer architecture was utilized as the primary framework. This model, proposed by Hong et al., leverages Transformer-based techniques to address the unique challenges of hyperspectral data, including high dimensionality, spectral redundancy, and spatial variability. The architecture combines spectral and spatial feature extraction through a patch-based representation and self-attention mechanisms, enabling the model to dynamically focus on the most relevant features. The incorporation of spectral embedding and position encoding ensures meaningful representations are learned while preserving spatial context and interdependence across spectral bands.

The SpectralFormer effectively handles noise and redundancy in hyperspectral data, improving classification accuracy even under challenging conditions. Its design enables scalability to large datasets, leveraging advanced Transformer techniques to extract nuanced spectral-spatial relationships. By integrating these capabilities, the SpectralFormer enhances the robustness and accuracy of HSI classification, demonstrating significant potential for applications in remote sensing, agriculture, and environmental monitoring. Figure 4 illustrates the SpectralFormer architecture, highlighting its modular components and its capacity to address the complexities of hyperspectral data. To enhance the model's robustness against potential adversarial attacks, we implemented an adversarial training strategy using the Fast Gradient Sign Method (FGSM).

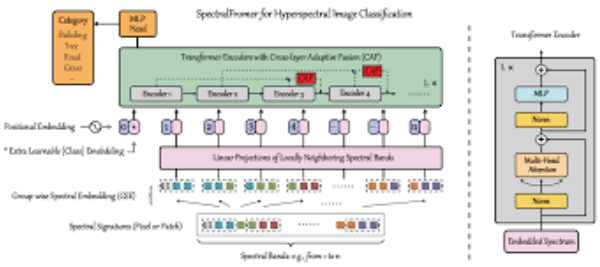

Fig. 3. The SpectralFormer architecture as proposed by Hong, et al.

B. Adversarial Training

The adversarial training methodology integrates the Fast Gradient Sign Method (FGSM) into the model optimization process through a weighted loss function:

$$\mathcal{L}_{total} = (1 - \alpha)\mathcal{L}_{standard} + \alpha\mathcal{L}_{adversarial}$$

where $\alpha \in [0, 1]$ controls the contribution of the adversarial component. During each training batch, FGSM generates adversarial examples by perturbing the input data according to:

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x \mathcal{L}(x, y))$$

where x represents the input data, and the step size ϵ constrains the magnitude of the perturbation. The model calculates both the standard cross-entropy loss on clean inputs and an adversarial loss on these perturbed samples. This formulation enables efficient training by generating adversarial examples only when $\alpha > 0$, allowing for a smooth transition between standard training ($\alpha = 0$) and adversarial training ($\alpha > 0$). By penalizing the model's sensitivity to small, targeted perturbations, the adversarial loss term enhances overall model robustness.

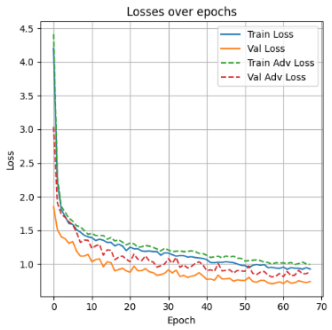

Fig. 4. The output of Training loss on the Pavia dataset at $\epsilon = .06$.

The training curves demonstrate the effects of FGSM adversarial training when evaluating model robustness. The validation loss displays pronounced oscillations because at each epoch, the validation set undergoes adversarial attacks, purposefully testing the model's resilience against perturbed inputs. This contrasts with the training phase, where the model learns from both clean and adversarial examples simultaneously. With an α value of 0.1, the model maintains a careful balance between standard performance and adversarial robustness, allowing us to systematically assess its defensive capabilities against potential attacks during validation.

C. Implementation Environment

All experiments were conducted on a single NVIDIA RTX 4070 GPU. The model implementation and training pipeline were developed using PyTorch, with all computations performed at 32-bit floating-point precision. This hardware configuration enabled efficient processing of the hyperspectral data and adversarial example generation during training.

IV. Experiments, Results and Discussion

A. Experimental Design

As outlined in the Datasets section, stratified sampling was employed to partition the data into Train, Validation, and Test sets. The model hyperparameters were configured with an alpha (α) value of 0.1 and a patch dimension of $7 \times 7 \times B$, where B represents the number of spectral bands. Training was conducted over 200 epochs using the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. Early stopping with a patience value of 30 epochs was implemented to prevent overfitting and ensure optimal model convergence.

To systematically evaluate the model's robustness across different perturbation magnitudes, we conducted experiments with ϵ values ranging from 0.01 to 0.15. The selected values ($0.01, 0.03, 0.06, 0.09,$ and 0.15) provided a comprehensive assessment of model performance under varying levels of adversarial attack strength. The upper bound of 0.15 was established as the threshold where perturbations became visually perceptible in the hyperspectral data, maintaining the requirement that adversarial modifications remain imperceptible to human observers.

For each epsilon value, we performed three independent training runs to ensure statistical reliability and account for training variability. The model's performance was evaluated using three key metrics: standard accuracy on clean test data, adversarial accuracy under FGSM attack, and an overall robustness score derived from the test set. These metrics were averaged across the three runs to obtain reliable performance estimates at each perturbation level. To establish a baseline for comparison, we additionally trained models using only standard loss at each epsilon value, enabling direct assessment of the adversarial training's impact on model robustness. This comprehensive experimental design enabled us to analyze the relationship between attack strength and model robustness while ensuring the reliability of our findings through multiple evaluations.

B. Results

We evaluated model performance across the three hyperspectral datasets using both standard and adversarial training methods. Our analysis measured clean accuracy (Clean Acc.), adversarial accuracy (Adv. Acc.), and robustness score (Rob. Score) under varying perturbation strengths ($\epsilon = 0.01$ to 0.15).

Our experiments revealed distinct patterns across the datasets. The Pavia dataset demonstrated exceptional resilience to adversarial attacks, maintaining robustness scores above 0.94 across all epsilon values. This dataset exhibited consistent clean accuracy (~ 0.99) regardless of training method and showed minimal difference between standard and adversarial training approaches. While adversarial accuracy gradually degraded from 0.989 to 0.941 as ϵ increased, Pavia maintained the highest overall robustness scores among all datasets.

The Salinas dataset, in contrast, showed greater sensitivity to adversarial attacks but responded well to adversarial training. Standard model robustness decreased substantially from 0.941

TABLE II
MODEL PERFORMANCE COMPARISON ACROSS DATASETS WITH DIFFERENT PERTURBATION STRENGTHS

Dataset	ϵ	**Clean Acc.**		**Adv. Acc.**		**Rob. Score**	
		Std	Adv	Std	Adv	Std	Adv
Pavia	0.01	0.990	0.992	0.987	0.989	0.997	0.997
	0.03	0.990	0.991	0.981	0.983	0.991	0.991
	0.06	0.991	0.991	0.968	0.972	0.977	0.979
	0.09	0.991	0.991	0.958	0.962	0.967	0.970
	0.15	0.991	0.991	0.939	0.941	0.947	0.949
Salinas	0.01	0.950	0.951	0.894	0.917	0.941	0.965
	0.03	0.962	0.950	0.802	0.853	0.834	0.898
	0.06	0.965	0.956	0.678	0.744	0.703	0.779
	0.09	0.959	0.949	0.601	0.690	0.626	0.726
	0.15	0.959	0.936	0.561	0.671	0.585	0.717
Indian Pines	0.01	0.979	0.975	0.950	0.942	0.971	0.965
	0.03	0.978	0.982	0.843	0.877	0.862	0.894
	0.06	0.979	0.982	0.684	0.794	0.699	0.809
	0.09	0.983	0.985	0.660	0.768	0.671	0.780
	0.15	0.983	0.981	0.575	0.734	0.599	0.747

Std: Standard training without adversarial loss; Adv: Training with adversarial loss
Results show averaged values over three trials for each configuration

($\epsilon = 0.01$) to 0.575 ($\epsilon = 0.15$). However, adversarial training improved robustness by 11% at $\epsilon = 0.15$, while clean accuracy remained consistently high (> 0.93) across all configurations. Among all datasets, Salinas demonstrated the most pronounced benefits from adversarial training.

The Indian Pines dataset exhibited patterns similar to Salinas, with notable improvements from adversarial training. While both training methods showed comparable performance at low ϵ (0.01), adversarial training yielded a 15% improvement in robustness at $\epsilon = 0.15$. The dataset maintained high clean accuracy (> 0.97) across all configurations, though standard model performance degraded significantly at higher ϵ values.

Cross-dataset analysis revealed several key patterns. Adversarial training proved most effective for Salinas and Indian Pines datasets, while showing minimal impact on Pavia's already robust performance. The effectiveness of adversarial training increased with ϵ across all datasets. Despite the varying levels of natural resilience, clean accuracy remained remarkably stable across all datasets, with Pavia consistently maintaining the highest robustness scores across all ϵ values.

These results demonstrate that dataset characteristics significantly influence both natural robustness and the potential benefits of adversarial training. While some datasets, like Pavia, exhibit inherent resilience to adversarial perturbations, others benefit substantially from defensive measures, particularly under stronger attacks.

C. Discussion

The experimental results demonstrate several significant findings regarding the robustness of SpectralFormer models in hyperspectral image classification. First, the SpectralFormer base architecture exhibits notable inherent robustness against adversarial attacks, particularly at lower perturbation strengths

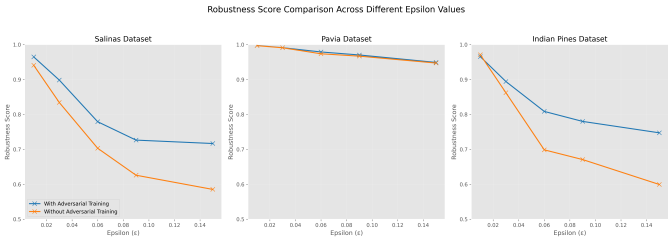

Fig. 5. Robustness score comparison across models.

($\epsilon \leq 0.03$). This suggests that the model's attention mechanisms and spectral feature extraction capabilities contribute to natural defensive properties against small perturbations.

A key observation is that adversarial training successfully enhanced model robustness without compromising clean accuracy. Although adversarial training often involves a trade-off between robustness and accuracy, our results show that clean accuracy remained stable across all datasets, with minimal degradation even under strong adversarial training regimes. This indicates that the SpectralFormer architecture can effectively incorporate adversarial defenses while maintaining its core classification capabilities.

The varying effectiveness of adversarial training across datasets provides insight into the relationship between data characteristics and model robustness. The Pavia dataset's minimal improvement under adversarial training, coupled with its high baseline robustness, suggests that certain spatial-spectral patterns may naturally resist adversarial perturbations. In contrast, the substantial improvements observed in the Salinas and Indian Pines datasets indicate that adversarial training can significantly improve robustness when natural resilience is lower.

Consistent performance improvements at higher epsilon values ($\epsilon \geq 0.06$) demonstrate that adversarial training par-

ticularly improves robustness against stronger attacks. This scalability of defense effectiveness suggests that the model learns generalizable robust features rather than becoming merely resistant to specific perturbation patterns.

V. Conclusion and Future Work

This study investigated the effectiveness of adversarial training in enhancing the robustness of SpectralFormer models for hyperspectral image classification. Our results demonstrated that while the SpectralFormer architecture exhibits natural resilience to adversarial attacks, particularly at lower perturbation strengths, adversarial training can further enhance this robustness without compromising clean accuracy. The effectiveness of this approach varied across datasets, with Pavia showing high inherent robustness, while Salinas and Indian Pines benefited significantly from adversarial training, especially under stronger attacks.

The successful implementation of adversarial defenses while maintaining classification performance suggests that transformer-based architectures can effectively balance robustness and accuracy in hyperspectral image classification tasks. Our findings indicate that the characteristics of the dataset play a crucial role in determining both natural resilience and the potential benefits of adversarial training, highlighting the importance of considering the properties of the data when developing robust classification systems.

Several promising directions for future work emerge from this study:

- **Average Accuracy:** Evaluating the model using average accuracy across classes to deal with examples of misbalanced classes in the ground truth data.
- **Architecture Optimization:** Investigation of simplified variants of the SpectralFormer architecture could reduce computational complexity while maintaining robustness benefits.
- **Hyperparameter Analysis:** A comprehensive study of hyperparameter optimization prior to adversarial testing could reveal optimal configurations for both performance and robustness.
- **Attack Diversity:** Evaluation against different attack methods, particularly Projected Gradient Descent (PGD), could provide broader insight into model robustness.
- **Alternative Architectures:** Comparative analysis with other state-of-the-art architectures, such as SACNet, could help identify the most effective approaches for robust hyperspectral image classification.

VI. Contributions

1) Joshua Orndorff:

- Developed all of the code from data-loading to model evaluation.
- Wrote the report.

2) Josue Syvelsaint:

- Created the visualizations of the project.
- Finished the slides for presentation.

VII. References

References

[1] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.

[2] R. Grewal, S. Singh Kasana, and G. Kasana, "Machine learning and deep learning techniques for spectral spatial classification of hyperspectral images: A comprehensive survey," *Electronics*, vol. 12, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/3/488

[3] S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231216310104

[4] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. M. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 2, pp. 6–36, 2013.

[5] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–15, 2022. [Online]. Available: http://dx.doi.org/10.1109/TGRS.2021.3130716

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2014. [Online]. Available: https://arxiv.org/abs/1312.6199

[7] Y. Xu, B. Du, and L. Zhang, "Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 2, pp. 1604–1617, 2021.

[8] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2574–2582.