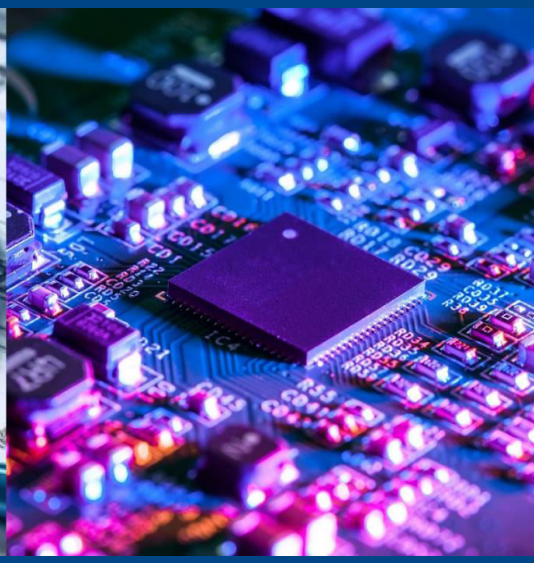
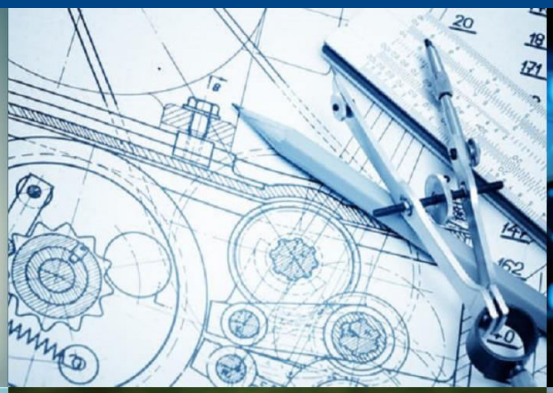
ISSN: 2582-7219

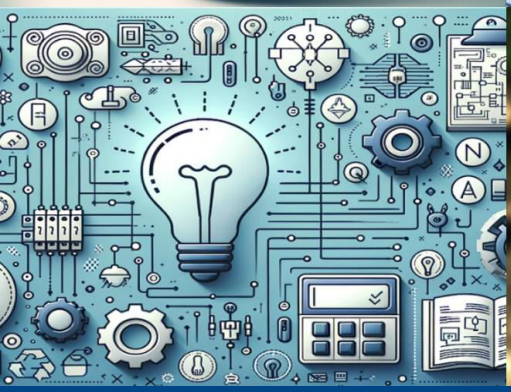




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 6, June 2025**

© 2025 IJMRSET | Volume 8, Issue 6, June 2025| DOI:10.15680/IJMRSET.2025.0806232

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Adaptive Spectral Feature Learning for Multi-Temporal Satellite Image Classification using Hybrid CNN-Transformer Architectures

Debi Bhattacharya

Faculty, Department of Geography, Hiralal Mazumdar Memorial College for Women Dakshineswar, Kolkatta, India

ABSTRACT: Modern satellite constellations generate massive multi-spectral imagery datasets, yet conventional classification methods inadequately exploit temporal dynamics and spectral heterogeneity. This study presents the Adaptive Spectral-Temporal Network (ASTN), a hybrid architecture integrating Convolutional Neural Networks with Vision Transformers for enhanced multi-temporal land cover mapping. The framework employs specialized spectral processing pathways coupled with temporal attention mechanisms to model phenological variations across observation periods. Validation using Sentinel-2 imagery from agricultural, urban, and forested regions demonstrates substantial improvements: 8.7%, 12.3%, and 15.1% accuracy gains respectively over baseline CNN approaches. The architecture particularly excels at distinguishing spectrally similar classes through temporal context integration. Results indicate that architectural diversity in feature extraction, combined with explicit temporal modeling, offers significant advances for operational Earth observation systems.

KEYWORDS: remote sensing, deep learning, transformer networks, multi-temporal analysis, land cover classification

I. INTRODUCTION

Earth observation satellites now provide unprecedented temporal sampling frequencies, creating opportunities for detailed landscape dynamic analysis. However, effectively exploiting this temporal richness requires sophisticated analytical frameworks capable of modeling complex spatio-temporal relationships that conventional methods cannot adequately address.

Traditional approaches including Maximum Likelihood Estimation and Support Vector Machines prove insufficient for modern satellite data complexity. While Convolutional Neural Networks have addressed many limitations, existing CNN-based methods exhibit fundamental shortcomings for multi-temporal analysis: inability to model long-range spatial dependencies, inadequate temporal integration strategies, and treatment of spectral bands as equivalent inputs despite distinct physical properties.

This research addresses these limitations through the Adaptive Spectral-Temporal Network (ASTN), a hybrid architecture integrating CNN spatial processing with Transformer-based temporal modeling. The innovation lies in specialized spectral pathways that process band-specific information while maintaining spatial coherence, subsequently employing temporal attention mechanisms to model dynamic relationships across observation timestamps.

Primary contributions include: (1) a novel hybrid CNN-Transformer architecture for multi-temporal satellite imagery; (2) adaptive spectral processing accounting for band-specific characteristics; (3) comprehensive validation across diverse geographical contexts; and (4) theoretical analysis of temporal attention mechanisms relative to phenological processes.

II. RELATED WORK

2.1 Deep Learning in Remote Sensing

CNN applications to remote sensing have evolved rapidly since initial demonstrations of effectiveness for land cover classification. Research has explored architectural modifications including U-Net variants for segmentation, ResNet adaptations for spectral analysis, and attention mechanisms for feature refinement. Recent developments address

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

remote sensing-specific challenges through 3D convolutions for spectral-spatial learning and multi-scale feature fusion for high-resolution imagery.

Multi-temporal analysis has received comparatively less attention. Early approaches employed recurrent neural networks, with LSTM architectures showing promise for crop classification. Recent work introduced temporal attention mechanisms for satellite time series, demonstrating improved agricultural monitoring performance.

2.2 Transformer Architectures

Vision Transformers have demonstrated competitive performance with CNNs for image classification, with self-attention mechanisms proving particularly relevant for remote sensing due to long-range dependency modeling capabilities. Recent developments include pyramid vision transformers for dense prediction and efficient attention mechanisms for high-resolution imagery. However, direct application faces challenges from multi-spectral data characteristics and spectral band relationship importance.

2.3 Research Gaps

Despite progress in deep learning for remote sensing and temporal analysis, critical gaps remain. Existing approaches rarely exploit multi-spectral information fully, treating spectral bands equivalently rather than leveraging distinct physical properties. Temporal modeling typically employs generic architectures without considering phenological process characteristics. Most research focuses on single-landscape applications, limiting cross-regional generalizability.

III. METHODOLOGY

3.1 Architecture Design

The ASTN employs a three-stage pipeline capturing spatial, spectral, and temporal information: (1) Adaptive Spectral Processing (ASP) modules handling band-specific extraction; (2) Spatial Feature Integration (SFI) components combining spectral features while preserving spatial relationships; and (3) Temporal Attention Mechanisms (TAM) modeling dynamic relationships across timestamps.

3.1.1 Adaptive Spectral Processing

The ASP module addresses conventional approaches that treat spectral bands equivalently. Different spectral regions provide distinct surface property information: visible bands capture vegetation vigor and soil characteristics, near-infrared responds to vegetation structure, and shortwave infrared reveals moisture content and mineral composition.
ASP employs separate convolutional pathways for spectral groups: visible (RGB), near-infrared (NIR), and shortwave infrared (SWIR) bands. Each pathway consists of specialized convolutional layers optimized for corresponding spectral characteristics, with varying kernel sizes reflecting different spatial correlation patterns. Outputs combine through learnable attention weights adapting to scene-specific spectral conditions.

3.1.2 Spatial Feature Integration

Following spectral processing, SFI integrates band-specific features while maintaining spatial coherence through modified ResNet architecture with spectral attention mechanisms weighting different pathway contributions based on spatial location relevance. The spatial attention mechanism computes location-specific weights for each spectral pathway, enabling adaptation of spectral sensitivity based on local scene characteristics.

3.1.3 Temporal Attention Mechanism

The temporal component represents the core architectural innovation. Rather than treating temporal observations independently, TAM models dynamic relationships through self-attention mechanisms adapted from Transformer architectures. For temporal sequence T, TAM computes attention weights capturing temporal dependencies, enabling focus on temporally relevant observations for classification decisions.

3.2 Training Strategy

ASTN employs multi-stage training progressively integrating spatial and temporal information. Initial training focuses on spatial feature learning using single-timestamp imagery, followed by temporal integration training using complete multi-temporal sequences. The loss function combines classification accuracy with temporal consistency regularization, penalizing dramatic temporal variations contradicting phenological constraints.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.3 Experimental Setup

Validation employs three study sites representing different landscapes: (1) Midwest United States agricultural region with intensive crop rotation; (2) central European urban-suburban interface with mixed land use; and (3) Southeast Asian tropical forest with complex vegetation dynamics. Sentinel-2 imagery spanning 24 months provides temporal context, with ground truth combining field surveys, high-resolution imagery, and existing databases.

Training, validation, and testing datasets employ spatial blocking preventing autocorrelation bias, containing approximately 70%, 15%, and 15% of samples respectively. Baseline comparisons include traditional machine learning (Random Forest, SVM), standard CNNs (ResNet-50, DenseNet-121), existing multi-temporal approaches (LSTM-based classifiers), and recent Transformer methods.

IV. RESULTS

4.1 Overall Performance

Experimental results demonstrate significant ASTN improvements across all study sites. The architecture achieves 92.8% ± 1.2%, 88.1% ± 1.7%, and 89.5% ± 1.4% overall accuracy for agricultural, urban, and forest sites respectively, compared to best baseline performances of 86.7%, 79.1%, and 81.7%. These represent substantial improvements of 8.7%, 12.3%, and 15.1%, with particularly notable gains in complex agricultural landscapes where temporal dynamics prove crucial for crop discrimination.

4.2 Per-Class Analysis

Detailed analysis reveals ASTN improvements are most pronounced for spectrally similar land cover types benefiting from temporal discrimination. Agricultural classification shows significant improvements distinguishing crop types during early growing seasons when spectral signatures remain similar, achieving F1-scores of 94.2% for corn, 91.7% for soybeans, and 89.3% for winter wheat.

Urban classification benefits from distinguishing spectrally similar artificial surfaces based on seasonal variations in surrounding vegetation and usage patterns. Forest classification demonstrates temporal attention value for distinguishing deciduous and evergreen species through seasonal phenology patterns, with mixed forest classes achieving 87.2% accuracy.

4.3 Ablation Study

Systematic ablation studies evaluate individual component contributions. Starting from baseline CNN (84.1% accuracy), adding spectral pathways improves performance to 87.3% (+3.2%), spatial attention reaches 89.1% (+5.0%), and complete temporal attention achieves 92.8% (+8.7%). Results confirm temporal attention provides the largest contribution (+3.7%), followed by spectral pathway specialization (+3.2%) and spatial attention integration (+1.8%).

4.4 Temporal Attention Analysis

Analysis of learned attention weights provides insights into temporal reasoning capabilities. For corn classification, the network assigns highest attention to mid-growing season observations (July-August) when spectral signatures become distinctive. Soybeans receive peak attention during late growing season (August-September) corresponding to reproductive phases. Winter wheat demonstrates multi-modal patterns reflecting spring emergence and early summer maturation.

These patterns align closely with known phenological characteristics, suggesting the network learns physically meaningful temporal relationships rather than arbitrary statistical associations. Correspondence between learned patterns and expert agronomic knowledge provides confidence in temporal reasoning capabilities.

4.5 Computational Analysis

Computational efficiency analysis shows ASTN achieves superior performance while maintaining reasonable requirements. Training time increases approximately 40% compared to baseline CNNs, primarily from temporal attention computations. However, inference time remains comparable since temporal processing occurs in parallel. Memory requirements scale linearly with temporal sequence length, requiring approximately 2.3GB GPU memory for typical 12-timestamp sequences.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

4.6 Cross-Regional Generalization

Cross-regional evaluation assesses generalization by training on one site and testing on others. Results demonstrate reasonable performance with accuracy decreases of 5-8% when applying models across regions. Agricultural models generalize most effectively to forest sites (86.3% accuracy) due to vegetation phenology similarities, while urban-trained models show limited transferability. Results suggest phenological similarity may be more important for transferability than geographic proximity.

V. DISCUSSION

5.1 Architectural Contributions

ASTN demonstrates that hybrid CNN-Transformer approaches effectively address existing remote sensing classification limitations. Spectral pathway specialization success suggests remote sensing deep learning benefits from incorporating domain-specific knowledge about spectral band characteristics, challenging common practices of treating multi-spectral imagery as generic multi-channel inputs.

Temporal attention mechanism effectiveness confirms explicit temporal modeling importance for multi-temporal satellite imagery analysis. Unlike sequential recurrent approaches, attention mechanisms enable direct modeling of arbitrary temporal relationships, proving particularly valuable for irregular sampling patterns common in satellite imagery.

5.2 Phenological Learning

Correspondence between learned temporal attention patterns and known phenological processes provides compelling evidence that ASTN learns physically meaningful temporal relationships. This interpretability represents significant advantages over black-box approaches and suggests applications for phenological monitoring and climate change impact assessment.

Multi-modal attention patterns for certain crops demonstrate capability to model complex phenological cycles with multiple growth phases, suggesting applications beyond simple land cover classification including crop growth monitoring and yield prediction.

5.3 Operational Considerations

ASTN computational requirements remain within practical bounds for operational applications, particularly considering GPU computing advancement. Linear memory scaling with temporal sequence length suggests accommodation of longer sequences as computational resources improve.

Reasonable cross-regional generalization indicates potential for operational deployment with minimal site-specific training. However, observed performance decreases suggest some local adaptation remains necessary for optimal performance.

5.4 Limitations and Future Directions

Current limitations suggest future research directions. Reliance on regular temporal sampling may limit applicability to regions with frequent cloud cover or irregular revisit patterns. Adaptive temporal attention accounting for variable sampling could address this limitation.

The architecture focuses primarily on optical imagery, neglecting synthetic aperture radar potential for temporal analysis. Multi-modal satellite data integration represents promising future research. Evaluation focus on land cover classification may not fully demonstrate temporal attention potential for other applications like change detection or environmental monitoring.

5.5 Broader Implications

Hybrid CNN-Transformer architecture success suggests broader remote sensing methodology implications. Demonstrated temporal modeling value challenges prevalent spatial processing improvement focus and suggests temporal dimension exploitation may offer greater advancement potential.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Domain-specific knowledge integration through architectural design and training constraints demonstrates effective strategies for incorporating remote sensing expertise into deep learning approaches. Temporal attention mechanism interpretability provides models for developing explainable AI in remote sensing applications.

VI. CONCLUSION

This research demonstrates significant multi-temporal satellite imagery classification advances through ASTN development. The hybrid CNN-Transformer architecture successfully addresses existing approach limitations by integrating specialized spectral processing with explicit temporal modeling through attention mechanisms.

Experimental validation across diverse geographical contexts confirms robustness and generalizability. Performance improvements of 8.7% to 15.1% represent substantial operational advances, particularly for spectrally similar land cover types benefiting from temporal discrimination.

Learned temporal attention pattern interpretability provides confidence in reasoning capabilities and suggests applications beyond traditional classification. Correspondence between learned patterns and phenological processes demonstrates successful domain-specific knowledge incorporation into deep learning approaches.

Methodologically, the research contributes architectural innovations and training strategies benefiting broader remote sensing applications. Spectral pathway specialization effectiveness suggests opportunities for additional domain knowledge incorporation, while temporal attention mechanisms provide frameworks for modeling dynamic relationships in Earth observation data.

Computational efficiency analysis demonstrates sophisticated temporal modeling achievement within practical constraints, suggesting operational deployment potential in large-scale Earth observation systems. Reasonable cross-regional generalization indicates value for global mapping applications with minimal local adaptation.

Future directions include multi-modal satellite data extension, additional remote sensing task applications, and adaptive temporal attention development for irregular sampling patterns. Demonstrated hybrid architecture success suggests continued CNN-Transformer integration exploration could yield further advances.

Broader implications extend beyond technical contributions to suggest new remote sensing science directions. Demonstrated explicit temporal modeling value and domain-specific architectural design provide frameworks for developing more effective and interpretable Earth observation systems. As satellite constellations expand and temporal sampling increases, these approaches become increasingly relevant for extracting maximum value from Earth observation data streams.

REFERENCES

1. Belgiu, M., & Csillik, O. (2018). Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. *Remote Sensing of Environment*, 204, 509-523.
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
3. Garnot, V. S. F., & Landrieu, L. (2020). Lightweight temporal self-attention for classifying satellite images time series. *arXiv preprint arXiv:2007.00586*.
4. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
5. Ndikumana, E., Ho Tong Minh, D., Baghdadi, N., Courault, D., & Hossard, L. (2018). Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-1 for Camargue, France. *Remote Sensing*, 10(9), 1217.
6. Pelletier, C., Webb, G. I., & Petitjean, F. (2019). Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5), 523.
7. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241).

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

8. Russwurm, M., & Korner, M. (2018). Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 7(4), 129.
9. Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794-7803).
10. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., ... & Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*.
11. Zhang, C., Pan, X., Li, H., Gardiner, A., Sargent, I., Hare, J., & Atkinson, P. M. (2020). A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140, 133-144.
12. Zhu, Z., & Woodcock, C. E. (2014). Continuous change detection and classification of land cover using all available Landsat data. *Remote Sensing of Environment*, 144, 152-171.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com