# AgroVision: Spatiotemporal Deep Learning for Large-Scale Crop Mapping and Food Security Assessment

Bill Gu
Webber Academy, Canada

Dr. Haoyan Jiang
FutureEra Research Institute, Canada

Project Track: Natural & Applied Science - Climate & Environment

Country / Region: Canada

June 26, 2026

***Abstract*—Ensuring global food security and promoting sustainable agriculture require timely, scalable, and reliable monitoring of crop systems, particularly in regions where ground-based data collection is limited. Advances in satellite remote sensing and artificial intelligence offer a promising pathway to address this challenge by enabling large-scale, data-driven agricultural analysis. In this work, we adopt a spatio-temporal deep learning framework for crop-type semantic segmentation using multi-temporal Sentinel-2 imagery. The approach integrates a U-Net architecture with convolutional Long Short-Term Memory units to jointly model spatial structure and seasonal crop dynamics. To enhance robustness and generalization, we augment the multispectral input with physically grounded vegetation and soil indices that encode agronomically meaningful biophysical properties. Collectively, this study highlights how combining spatio-temporal modeling, domain-informed representation learning, and open big data can contribute to scalable agricultural intelligence in support of food security and sustainable agriculture. Project Track: Climate & Environment**

## I. Introduction

Crop failure, regardless of cause, is a great inhibitor of food security. To ensure sustainability, it is imperative that crop health is monitored to minimize yield loss. Unfortunately, long-term environmental changes, such as global warming, can heavily affect all environments, requiring farmers to constantly adapt to new conditions. Additionally, these methods require extra resources from farmers, limiting their uses in lower socioeconomic status areas. Satellite imagery, on the other hand, is both more accessible and provides great insight into crop health and failure predictions, allowing farmers to plan accordingly.

Modern developments in satellite technology have numerous advantages compared to crop analysis. Higher-resolution satellite images improve crop yield predictions, which emphasizes on the importance of quality satellite images. [1] Luckily, large, publicly available datasets from satellite imagery and land surface analysis are both accessible and abundant, utilizing strong foundational methods such as random forest and neural networks.

In this work, we propose a spatio-temporal crop-type segmentation framework that integrates convolutional and recurrent modeling with physically grounded spectral enrichment. Building upon a U-Net architecture augmented with convolutional Long Short-Term Memory units, the proposed approach captures spatial structure and seasonal crop dynamics from multi-temporal Sentinel-2 imagery. In addition to raw spectral bands, we incorporate a set of vegetation and soil indices as additional input channels, embedding agronomic priors directly into the learning process.

The contributions of this work are threefold: (1) a spatio-temporal deep learning framework for crop-type semantic segmentation that integrates spatial and phenological information, (2) a physically grounded spectral augmentation strategy that improves robustness and generalization, and (3) an end-to-end, reproducible pipeline that highlights the potential for scalable deployment in support of global food security and sustainable agriculture.

### *A. Spectral Indices and Domain Knowledge Integration*

Vegetation indices such as NDVI, red-edge indices, and moisture-related metrics have long been used to characterize vegetation vigour, chlorophyll concentration, and water stress [2, 3, 4]. While deep learning models are theoretically capable of learning such relationships implicitly, several studies report improved stability and interpretability when physically meaningful indices are explicitly included as input features [5, 6].

This hybrid strategy—combining raw spectral bands with engineered indices—has proven particularly effective in agricultural remote sensing, where biophysical interpretability and robustness across regions and seasons are critical. Such approaches align well with sustainability-oriented applications, as they are both transparent and consistently accurate.

## II. Methods

### *A. Dataset*

This study is conducted using the *PASTIS* dataset[7], a broad, open-source benchmark specifically designed for panoptic segmentation of agricultural landscapes from multi-temporal Sentinel-2 imagery [8]. PASTIS provides a unique combination of dense pixel-level semantic labels and instance-level parcel annotations, making it particularly well-suited for learning both crop-type semantics and field-level spatial structure.

The dataset is composed of multi-date Sentinel-2 image time series covering diverse agricultural regions across France. Each sample consists of a sequence of atmospherically corrected surface reflectance observations acquired throughout the growing season, enabling the model to capture phenological dynamics rather than relying on single-date spectral snapshots. Temporal sampling is irregular but sufficiently dense to reflect key crop development stages, which is critical for discriminating between visually similar crop types.

Spatially, PASTIS integrates two complementary annotation modalities. At the pixel level, semantic segmentation labels assign each pixel to a crop category or background class. At the instance level, vector-based parcel geometries delineate individual agricultural fields, allowing panoptic supervision that jointly models *what* crop is present and *where* each field instance begins and ends. This dual supervision paradigm is particularly advantageous for precision agriculture tasks, where field boundaries and crop identity are both essential outputs.

All imagery is derived from Sentinel-2 Level-2A products and includes multiple spectral bands spanning the visible, red-edge, near-infrared (NIR), and shortwave infrared (SWIR) regions, providing multiple perspectives of analysis.

### B. Spectral and Index-Based Data Augmentation

Rather than relying solely on conventional geometric transformations, this work adopts a physically grounded data augmentation strategy based on spectral enrichment and index engineering. Given the radiometric consistency of Sentinel-2 imagery and the agronomic interpretability of vegetation indices, augmenting the input space through additional spectral channels provides a principled mechanism to expose latent biophysical signals and improve model generalization.

*a) Multispectral Channel Utilization:* Each Sentinel-2 observation inherently contains complementary spectral responses associated with vegetation structure, chlorophyll concentration, canopy water content, and soil exposure. In this study, the original input channels include visible bands (B2, B3, B4 at 10 m resolution), red-edge bands (B5, B6, B7 at 20 m), near-infrared bands (B8 at 10 m and B8A at 20 m), and shortwave infrared bands (B11 and B12 at 20 m). These bands collectively span key portions of the electromagnetic spectrum relevant to crop discrimination, particularly for separating phenologically similar classes that diverge in canopy density or moisture stress.

To ensure spatial consistency across channels, all 20 m bands are resampled to 10 m resolution prior to model ingestion. This harmonization enables pixel-aligned fusion while preserving the relative spectral relationships essential for index computation.

*b) Engineered Spectral Indices.:* In addition to the raw multispectral bands, several spectral indices are computed and concatenated to the input tensor to enhance sensitivity to vegetation structure, chlorophyll concentration, soil exposure, and moisture content. These indices are derived from established radiometric relationships and serve as compact representations of biophysical properties that are otherwise distributed across multiple spectral bands. (See Appendix A)

Collectively, these engineered indices enrich the spectral representation by embedding agronomic priors directly into the learning process.

### C. Model Architecture

*a) Spatial Modeling with U-Net.:* Following the approach proposed by Rustowicz et al. [9] for crop-type semantic segmentation, U-Net [10] serves as the core spatial feature extractor in the spatial-temporal modal. Its encoder–decoder architecture progressively aggregates contextual information through hierarchical convolution and downsampling, while skip connections directly propagate high-resolution spatial features to the decoder. This design enables precise pixel-level localization while retaining semantic context, a property that is particularly important in agricultural imagery where accurate delineation of field boundaries and within-field consistency are critical. By operating fully convolutionally, U-Net further supports dense prediction over large spatial extents without requiring patch-wise classification.

*b) Temporal Dependency Modeling with ConvLSTM.:* To incorporate temporal information inherent in the Sentinel-2 time series, ConvLSTM layers [11] are integrated into the network. ConvLSTM extends the classical LSTM formulation by replacing fully connected operations with convolutional kernels, thereby preserving spatial structure while modeling temporal evolution. Given a sequence of feature maps $\{\mathbf{X}_t\}_{t=1}^{T}$ extracted from consecutive observations, ConvLSTM learns hidden states $\mathbf{H}_t$ that encode phenological progression and temporal dependencies. This capability is essential for crop-type discrimination, as many crops exhibit similar spectral characteristics at isolated timestamps but diverge significantly when observed across the growing season.

*c) U-Net + ConvLSTM Integration.:* Following the design introduced by Rustowicz et al. [9], ConvLSTM units are embedded within the U-Net framework to

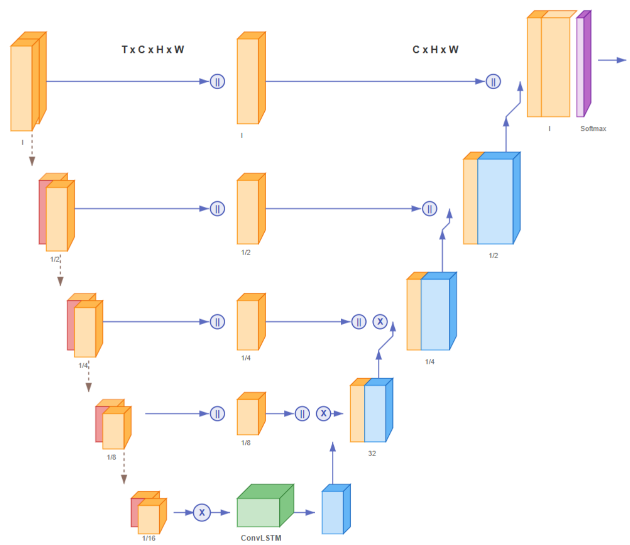


Figure 1. Unet - CovLSTM architecture.

operate on intermediate feature representations rather than raw pixel values as shown in figure 1. This integration allows temporal modeling to occur at a semantically meaningful level, where spatial patterns corresponding to crop parcels have already been abstracted. The resulting architecture jointly learns spatial features and their temporal transitions, enabling the model to distinguish crops based on both appearance and growth dynamics.

*d) Role of Spectral and Index-Based Channel Enrichment.:* The augmented spectral representation described in Section II-B is incorporated directly at the network input by concatenating engineered vegetation and soil indices with the original Sentinel-2 bands. This expanded input tensor provides the convolutional encoder with access to physically interpretable biophysical cues, such as vegetation vigor, chlorophyll concentration, moisture status, and soil exposure. By embedding these domain-informed signals into the learning process, the model is less reliant on discovering such relationships implicitly, leading to improved robustness under spectral noise, phenological variability, and heterogeneous field conditions.

The enriched feature space further benefits temporal modeling, as ConvLSTM units operate on representations that already encode agronomically meaningful structure. This facilitates more stable sequence learning and enhances generalization across crop types and geographic regions.

*e) Prediction Objective.:* The network produces dense pixel-wise predictions corresponding to crop-type semantic labels. Training is performed end-to-end using supervised learning, optimizing a categorical segmentation loss over all spatial locations and temporal observations. Through the combination of U-Net's spatial precision, ConvLSTM's temporal awareness, and physically grounded spectral augmentation, the proposed framework is well aligned with the demands of crop-type segmentation in multi-temporal satellite imagery.

### *D. Training and Deployment Pipeline*

The model was trained for approximately four hours on an NVIDIA Tesla P100 GPU via Kaggle, with atmospherically corrected Sentinel-2 Level-2A imagery and the PASTIS dataset.

Following training, the learned model weights are exported and integrated into an interactive inference pipeline. The trained model is deployed to HuggingFace Spaces, enabling real-time semantic segmentation inference and visualization through a lightweight web-based interface. This deployment facilitates qualitative evaluation, demonstration of temporal consistency in predictions, and broader accessibility for external users without requiring specialized hardware or local environment setup.

### *E. Overall Research Workflow and Implementation Timeline*

The overall research workflow consisted of four sequential stages: dataset acquisition, data preprocessing, model construction, experimental training, and deployment-based validation. In the first stage, the public PASTIS benchmark was selected as the primary data source because it provides multi-temporal Sentinel-2 image sequences together with dense crop-type annotations. In the second stage, the multispectral image sequences were harmonized by aligning all spectral bands to a common spatial resolution, normalizing reflectance values, and constructing additional vegetation and soil indices. In the third stage, a U-Net and ConvLSTM-based architecture was implemented to jointly capture spatial field structure and temporal crop development. In the fourth stage, the model was trained and evaluated using supervised semantic segmentation, with performance assessed through loss curves, mean Intersection-over-Union (mIoU), and ablation comparison.

The implementation timeline followed the same logic. Dataset preparation and preprocessing were completed first, followed by model implementation and debugging. Model training was then conducted on a Kaggle NVIDIA Tesla P100 GPU for approximately four hours. After training, validation metrics and visual outputs were examined to assess whether the learned model produced spatially coherent crop segmentation maps. The final deployment stage translated the trained model into a lightweight web-accessible inference system, enabling users to inspect model predictions without requiring local GPU resources.

## III. Results

This section presents the quantitative and qualitative evaluation of the proposed U-Net+ConvLSTM framework, including training dynamics, test-set performance, and an ablation analysis assessing the contribution of spectral and index-based channel enrichment.

### A. Training Dynamics

Figure 2 illustrates the evolution of training and validation performance across epochs. The mean Intersection-over-Union (mIoU) increases steadily for both training and validation sets during the early stages of optimization, indicating effective learning of spatial–temporal crop representations. While the training mIoU continues to improve throughout the training process, the validation mIoU exhibits a gradual saturation after approximately 15 epochs, stabilizing around a value of 0.47. This behavior suggests that the model reaches a point of diminishing returns in generalization while continuing to refine dataset-specific representations.

A similar pattern is observed in the loss curves. Both training and validation losses decrease rapidly during the initial epochs, reflecting fast convergence. As training progresses, the validation loss plateaus while the training loss continues to decrease, indicating mild overfitting. However, the absence of divergence between training and validation loss suggests that overfitting remains controlled and does not severely compromise generalization.

### B. Test Set Performance

Evaluation on the held-out test set confirms the generalization capability of the proposed approach. The model achieves a test loss of 0.0426 and a test mean Intersection-over-Union of 0.4729. These results are consistent with the validation performance observed during training, indicating that the learned spatio-temporal representations transfer effectively to unseen data. Given the inherent difficulty of crop-type segmentation—where inter-class spectral similarity and intra-class phenological variability are common—this level of performance demonstrates the suitability of the U-Net+ConvLSTM architecture for multi-temporal agricultural remote sensing.

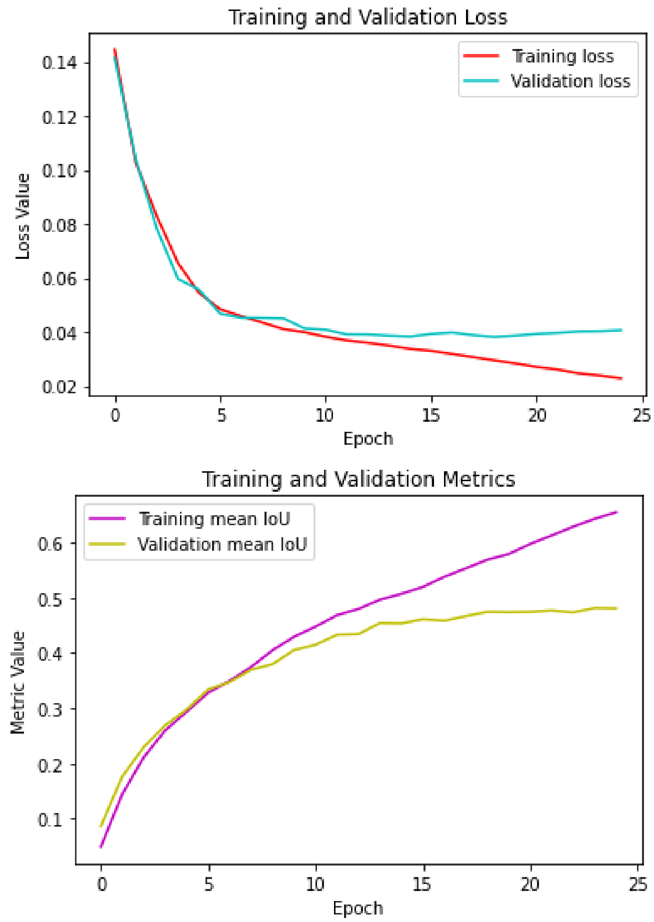


Figure 2. Training and validation performance over epochs. Top: segmentation loss. Bottom: mean Intersection-over-Union (mIoU).

### C. Ablation Study

To assess the impact of spectral and index-based channel enrichment, an ablation study is conducted comparing the full model against a baseline variant trained using only the original Sentinel-2 spectral bands. The baseline model achieves a test loss of 0.0438 and a test mIoU of 0.4543, which is consistently lower than the performance obtained with the augmented input representation.

Although the absolute improvement in mIoU is modest, the observed gain is systematic and aligns with the intended role of the engineered channels. By

Table I
TEST-SET PERFORMANCE OF THE BASELINE AND PROPOSED MODELS.

| Model | Test Loss | Test mIoU |
| --- | --- | --- |
| Sentinel-2 bands only | 0.0438 | 0.4543 |
| Sentinel-2 bands + spectral indices | 0.0426 | 0.4729 |

incorporating vegetation and soil indices that encode biophysical priors, the enriched model benefits from enhanced sensitivity to vegetation vigor, chlorophyll content, moisture status, and soil exposure. These factors are particularly important for distinguishing crop types with similar spectral signatures at individual time steps but divergent phenological trajectories.

## IV. DISCUSSION

### A. Implications for Food Security and Sustainable Agriculture

Besides our modeling emphasing on panoptics and crop-type segmentation, the main objective of this research was to explore the possiblility to build upon the massively available set of public satellite data and creating a agricultural monitoring tool at a scale using U-net and ConvLSTM that is accessible to farmers of all areas with available satellite. Even with a simple 3-6 months forecast, agricultural predictions greatly assist farmers to plan for future forecasting to various degrees [12].

The training and validation curves further suggest that the proposed model learns spatio-temporal representations from the satellite image time series. Training loss decreases steadily across epochs, while mean Intersection-over-Union improves consistently during the early and middle stages of optimization. Validation performance begins to plateau after approximately 25 epochs, indicating that the model approaches its generalization capacity under the current architecture and dataset configuration. The final validation performance, which approaches an mIoU of 0.5, suggests that temporal feature learning contributes useful information for crop identification in multi-date imagery. This pattern supports the broader premise that crop classes are more effectively distinguished when their seasonal development trajectories are modeled rather than when individual satellite observations are treated independently.

Beyond crop identification, the spectral and index-based enrichment strategy adopted in this work aligns closely with sustainability objectives. Indices related to vegetation vigor, chlorophyll content, moisture status, and soil exposure provide indirect indicators of crop health and land management practices, encouraging sustainable agricultural practices for farmers of various global backgrounds.

Moreover, the usage of the freely available Sentinel-2 imagery and open-source datasets such as PASTIS underscores the accessibility of the approach. This lowers barriers to adoption for regions with limited technological infrastructure, promoting more equitable access to AI-driven agricultural insights and supporting sustainable development goals.

### B. Limitations and Broader Considerations

While the results are encouraging, it is important to contextualize them within the broader challenges of sustainable agriculture. Crop-type segmentation alone does not directly address yield variability, market access, or socio-economic constraints faced by farmers. Additionally, optical satellite data remain sensitive to cloud cover and atmospheric conditions, which may limit applicability in certain regions or seasons. There is also limited recommended actions provided for farmers facing potential crop failures from the predictions, meaning that there is room to improve and potentially branch off into other useful models to assist farmers of different experiences and global backgrounds.

Overall, even with some of the restrictions due to the nature of the study, it still illustrates how advances in spatial-temporal deep learning, when coupled with open big data resources and accessible deployment pipelines, can support the broader goal of strengthening food security and promoting sustainable agriculture through scalable, data-driven insight.

## V. CONCLUSIONS

This study presents a spatial-temporal deep learning framework for crop-type semantic segmentation from multi-temporal Sentinel-2 imagery, motivated by the broader objective of leveraging big data and artificial intelligence to strengthen food security and promote sustainable agriculture. Integrating the U-Net architecture with ConvLSTM-based temporal modeling and enriching the input space with physically grounded spectral indices, the proposed approach demonstrates stable convergence, competitive generalization performance, and measurable gains in segmentation accuracy.

Beyond methodological performance, the implications of this work extend toward scalable agricultural intelligence in data-scarce and food-intensive regions. The reliance on freely available satellite imagery and open-source tooling positions the framework as a viable component of large-scale crop monitoring systems, particularly in regions such as sub-Saharan Africa where ground-based surveys are costly, sparse, or temporally delayed. Accurate crop-type mapping at scale can inform early-season production estimates, identify shifts in cropping patterns, and support evidence-based policy interventions aimed at mitigating food insecurity.

From a sustainability perspective, the incorporation of vegetation and soil indices enables learning from biophysical signals directly linked to crop health, moisture status, and land management practices. This comes at much uses in regions of limited technological developments.

Future research directions include extending the framework to panoptic segmentation objectives, incorporating complementary data sources such as synthetic aperture radar (SAR) to improve robustness under adverse imaging conditions, and scaling inference to a continental level [13]. Integrating crop segmentation outputs with yield modeling, market data, and socio-economic indicators also presents a promising avenue for translating technical advances into tangible societal impact. Collectively, this work highlights how thoughtfully designed AI systems, grounded in domain knowledge and supported by open big data, can contribute to more resilient, equitable, and sustainable global food systems.

## References


[1] Dhahi Al-Shammari, Brett M Whelan, Chen Wang, Robert GV Bramley, Mario Fajardo, and Thomas FA Bishop. Impact of spatial resolution on the quality of crop yield predictions for site-specific crop management. *Agricultural and Forest Meteorology*, 310:108622, 2021.

[2] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering. Monitoring vegetation systems in the great plains with erts. *NASA Special Publication*, 351:309–317, 1974.

[3] Anatoly A. Gitelson and Mikhail N. Merzlyak. Remote estimation of chlorophyll content in higher plant leaves. *International Journal of Remote Sensing*, 24(14):2917–2933, 2003.

[4] Bo-Cai Gao. Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3):257–266, 1996.

[5] Aaron E. Maxwell, Timothy A. Warner, and Fang Fang. Implementing machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, 39(9):2784–2817, 2018.

[6] Marc Rußwurm and Marco Körner. End-to-end learning for early classification of time series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:471–483, 2020.

[7] Vivien Sainte Fare Garnot. Pastis benchmark: Panoptic agricultural satellite time series. https://github.com/VSainteuf/pastis-benchmark, 2021. Accessed: 2026-01-09.

[8] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. *ICCV*, 2021.

[9] Rose Rustowicz, Robin Cheong, Luyu Wang, Stefano Ermon, Marshall Burke, and David Lobell. Semantic segmentation of crop type in africa: A novel dataset and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[11] Xingjian Shi, Zhihan Chalsen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-Chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems (NeurIPS)*, 28, 2015.

[12] James W Jones, James W Hansen, Frederick S Royce, and Carlos D Messina. Potential benefits of climate forecasting to agriculture. *Agriculture, ecosystems & environment*, 82(1-3):169–184, 2000.

[13] Giorgio Franceschetti and Riccardo Lanari. *Synthetic aperture radar processing*. CRC press,

2018.

[14] Akira Rikimaru, P. S. Roy, and S. Miyatake. Estimation of soil erosion using remote sensing. *International Journal of Remote Sensing*, 2002.

[15] Carl H. Key and Nathan C. Benson. Landscape assessment: Ground measure of severity, the composite burn index. *USDA Forest Service General Technical Report*, 2006.

[16] Alfredo R. Huete. A soil-adjusted vegetation index (savi). *Remote Sensing of Environment*, 25(3):295–309, 1988.

# Appendix A

Table II
SPECTRAL AND INDEX-BASED CHANNEL ENRICHMENT USED IN THE PROPOSED FRAMEWORK.

| **Index** | **Definition** | **Physical Interpretation and Modeling Role** |
|---|---|---|
| Normalized Difference Vegetation Index (NDVI) | $\frac{B8-B4}{B8+B4}$ | Vegetation vigor and green biomass; improves early-stage crop discrimination and temporal stability. [2] |
| Red-Edge Normalized Difference Index (NDRE5) | $\frac{B8-B5}{B8+B5}$ | Chlorophyll sensitivity in dense canopies where NDVI saturates; improves mid/late-season separability. [3] |
| Normalized Difference Water In- dex (NDWI) | $\frac{B8-B11}{B8+B11}$ | Canopy/surface moisture; improves robustness to irrigation and precipitation variability. [4] |
| Bare Soil Index (BSI) | $\frac{(B11+B4)-(B8+B2)}{(B11+B4)+(B8+B2)}$ | Bare soil exposure; separates fallow fields and sparse vegetation. [14] |
| Normalized Burn Ratio (NBR) | $\frac{B8-B12}{B8+B12}$ | Canopy stress / disturbance; improves generalization under heterogeneous conditions. [15] |
| Soil-Adjusted Vegetation Index (SAVI) ($L = 0.5$) | $\frac{(B8-B4)(1+L)}{B8+B4+L}$ | Soil-adjusted vegetation response; stabilizes early phenological representations. [16] |