engineering proceedings

MDPI

Proceeding Paper

Satellite-Based Crop Recognition Using Virtual Transformer Models for Smart Agriculture †

Kusum Lata [1,*], Navneet Kaur [1] and Simrandeep Singh [2]

[1] Department of Computer Science Engineering, Chandigarh University, Mohali, India; email1@email.com
[2] Department of Electronics & Communication Engineering, UCRD, Chandigarh University; email2@email.com
* Correspondence: verma.kusum91@gmail.com
† Presented at the 12th International Electronic Conference on Sensors and Applications (ECSA-12), 12–14 November 2025; Available online: https://sciforum.net/event/ECSA-12.

Abstract

Precision agriculture is dependent on precise crop identification to maximize resource utilization and enhance yield forecasting. This paper investigates the use of Vision Transformers (ViTs) for crop classification from high-resolution satellite images. In contrast to traditional deep learning models, ViTs use self-attention mechanisms to capture intricate spatial relationships and improve feature representation. The envisioned framework combines preprocessed multispectral satellite imagery with a Vision Transformer model that is optimized to classify heterogeneous crop types more accurately. Experimental outcomes confirm that ViTs are superior to conventional Convolutional Neural Networks (CNNs) in processing big agricultural datasets, yielding better classification accuracy. The proposed model was tested on a multispectral satellite image from Sentinel-2 and Landsat-8. The results shows that ViTs efficiently captured long-range dependencies and intricate spatial patterns and attained a high classification accuracy of 94.6% and a Cohen's kappa coefficient of 0.91. The incorporation of multispectral characteristics like NDVI and EVI also improved model performance, allowing for improved discrimination between crops with comparable spectral signatures. The results point out the applicability of Vision Transformers in remote sensing for sustainable and data-centric precision agriculture. Even with the improvements made in this study, issues like high computational expense, data annotation needs, and environmental fluctuations are still major hurdles to widespread deployment.

Keywords:

Academic Editor(s): Name

Published: date

Citation: Lata, K.; Kaur, N.; Singh, S. Satellite-Based Crop Recognition Using Virtual Transformer Models for Smart Agriculture. *Eng. Proc.* **2025**, *volume number*, x. https://doi.org/10.3390/xxxxx

1. Introduction

Precision agriculture has emerged as a transformative approach in modern farming, leveraging advanced technologies to optimize resource utilization, enhance crop yield, and ensure sustainable agricultural practices [1]. A fundamental aspect of precision agriculture is accurate crop identification, which aids in monitoring crop health, predicting yields, and implementing data-driven decision-making processes. Conventional crop classification approaches are based on field surveys by hand or traditional machine learning algorithms, which tend to lack scalability and accuracy [2]. The combination of satellite imaging with sophisticated deep learning methods holds the key to effective and high-

accuracy crop identification [3]. Figure 1 shows the important keywords used in remote sensing.

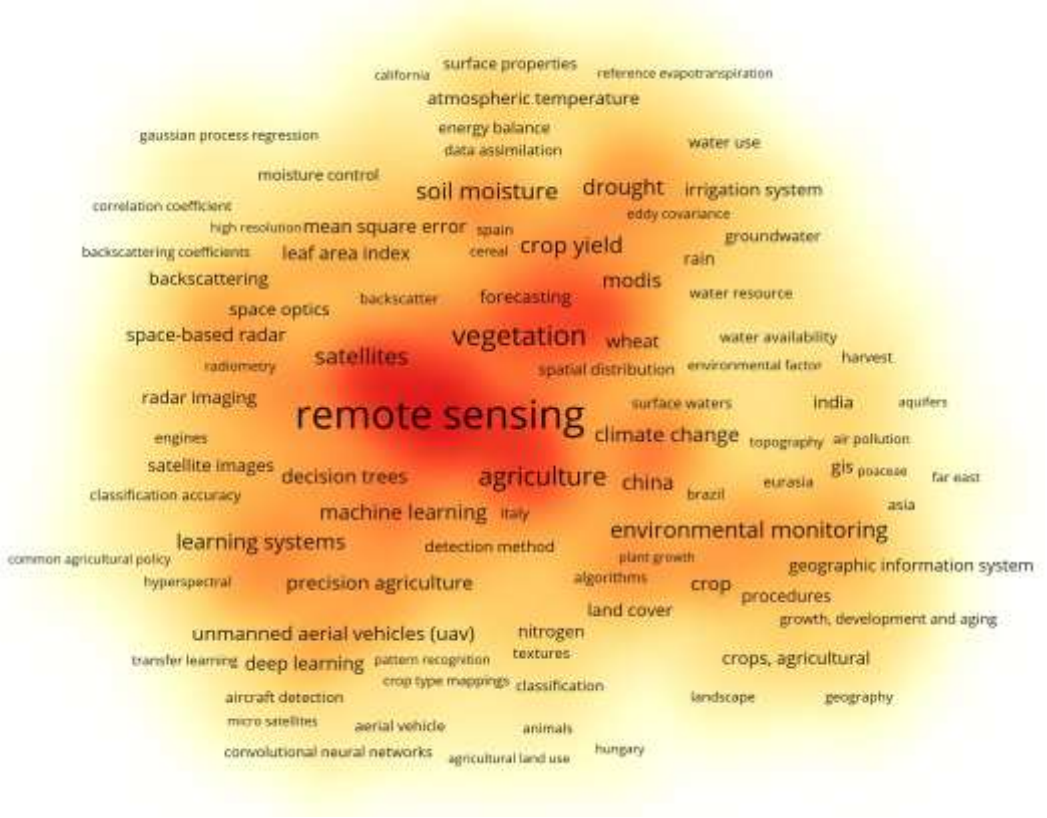

Figure 1. Some Important Keywords used in Remote Sensing.

New developments in computer vision, specifically deep learning architectures, have dramatically enhanced the pre-cision of agricultural remote sensing applications [4]. Convolutional Neural Networks (CNNs) have already been applied extensively to crop classification, but their limited receptive field and susceptibility to long-range dependencies limit their performance in heavy-tailed agricultural scenes [5]. To address these issues, Vision Transformers (ViTs) have proven to be a strong contender, with enhanced spatial feature ex- traction and better representation learning capabilities. Vision Transformers utilize self-attention to process global relations between image elements and achieve top performance in satellite image analysis [6]. The capabilities of ViTs exceed local areas of attention because they excel at learning complicated spatial structures needed for crop type discrimination. The ability to process large agricultural data sets while gaining contextual knowledge improves crop classification precision which leads to more dependable precision agriculture solutions [7]. Satellite imagery analyzed with ViTs produces superior outcomes than regular procedures because of multiple advantages. Big agricultural areas can portray their expansive crop status through real-time satellite imaging which delivers broad viewing capabilities [8]. This paper demonstrates how the combination of multispectral and hyperspectral data enables the method to extract vital vegetation indices and spectral signatures which identify different crop species. Real-time Use of these numerous data sources with ViTs produces better crop classification systems that demonstrate enhanced stability. This paper presents a framework built on Vision Transformers which performs satellite image-based crop classification [2]. The high-resolution agricultural datasets need multiple spectral bands because this training enables better classification accuracy. The performance evaluation of ViTs and CNN-based architectures happens through extensive testing experiments designed to examine their effectiveness in real

farming situations. The proposed system aids the progress of intelligent agricultural techniques through its delivery of accurate crop area identification capabilities to farmers and policy makers [9]. Although Vision Transformers (ViTs) have been extensively applied in remote sensing, including multispectral and hyperspectral imagery, most studies focus on generic land-use or limited crop types. This work introduces a novel ViT-based framework specifically designed for crop recognition in precision agriculture. By integrating multispectral features with vegetation indices (NDVI, EVI), the model enhances discrimination among spectrally similar crops and achieves higher accuracy than CNN-based approaches. The validation on real agricultural datasets highlights its robustness and contribution to sustainable crop monitoring.

2. Literature Review

Remote sensing and machine learning in crop classification are points of big research, however, a few problems will still exist. Bargiel [11] described an approach that used radar time-series and crop phenology together, but the single use of SAR data rendered the method less universally applicable compared to those surveyed in this article. Panigrahi et al. [12] compared several supervised ML regression models (M5 model tree and gradient-enhanced tree) through crop yield prediction which achieved good performance but showed that they could not easily handle the high-dimensional multispectral data. Mohanty et al. [13] have compared the ML approaches to the remote sensing yield prediction; however, their traditional ML models were not able to reflect the nonlinearity in large-scale data.

DNN have been known to be beneficial in predicting crops. Artificial neural networks were used by Shankar et al. [14] to estimate the effect of nutrients on the growth of rice, since they obtained better results than linear models, but often required a lot of tuning and labeled data. You et al. [15] introduced deep Gaussian processes to the crop yield prediction utilizing satellite imagery where it showed an improvement in prediction with reduced interpretability and scalability. More recently, deep learning has been used to count plants in aerial imagery [16]. Unlike multi-class crop recognition, this application is shown to be robust to plant occlusion.

There has also been adoption of vision-based deep learning including architectures. Kussul et al. [17] applied satellite images and deep neural nets to crop classification in Ukraine, providing another example of applying the power of big data but dealing with complexities of computation. Ji et al. [18] combined multi-temporal Sentinel-2 with recurrent neural networks (RNNs) to map crops, performing better than random forests and unable to generalise across seasons. Other Sentinel-2 use-cases Temporal convolutional networks were used by Russwurm and Kormer [19] to classify crop types and achieved improvements in capturing seasonal dynamics, but also needed dense time-series data.

In spite of these developments, there are still major gaps. Available solutions tend to use one type of data (SAR, optical, or UAV) and fail to perceive spectrally similar crops. Most ML/DL techniques need access to large and labeled data and struggle with scaling to heterogenous farming areas. In this respect, ViTs provide a significant development opportunity to improve long-range dependencies and sophisticated spatial patterns using self-attention. In contrast to CNN- or RNN-based methods, we combine ViT and multispectral data and vegetation indices (NDVI, EVI) the results of which appear to be more separable among spectrally mutually overlapping crops and better in accuracy on real-world agricultural data.

Table 1 shows the summary of the literature.

Table 1. Summary of references with key findings and research gaps.

Ref No.	Author(s) & Year	Title	Findings	Research Gaps
[11]	D. Bargiel, 2017	A new method for crop classification combining time series of radar images and crop phenology information	Demonstrated crop classification using radar time series with phenological information	Limited transferability; depends heavily on SAR data only
[12]	B. Panigrahi et al., 2022	A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models	Compared ML regression models for yield prediction with promising results	Models struggled with high-dimensional multispectral data
[13]	R. K. Mohanty et al., 2022	Comparative analysis of machine learning techniques for crop yield prediction using remote sensing data	ML models effective for yield estimation with remote sensing inputs	Traditional ML approaches failed to capture nonlinear relationships
[14]	T. Shankar et al., 2022	Prediction of the effect of nutrients on plant parameters of rice by artificial neural network	ANN outperformed linear models for rice growth prediction	Required extensive tuning and labeled data
[15]	J. You et al., 2017	Deep Gaussian process for crop yield prediction based on remote sensing data	Improved crop yield prediction using deep Gaussian processes	Limited interpretability and scalability in classification tasks
[16]	M. Rahnemoonfar and C. Sheppard, 2017	Deep count: Fruit counting based on deep simulated learning	Showed robustness of deep learning for plant/fruit detection even under occlusion	Did not extend to multi-class crop recognition
[17]	N. Kussul et al., 2017	Deep learning classification of land cover and crop types using remote sensing data	Demonstrated effectiveness of deep neural networks for crop classification in Ukraine	Faced issues of computational complexity and scalability
[18]	S. Ji et al., 2018	3D convolutional neural networks for crop classification with multi-temporal remote sensing images	RNN/3D CNN models improved classification from multi-temporal data	Poor generalization across different seasons/regions
[19]	M. Russwurm and M. Körner, 2019	Temporal convolutional neural networks for the classification of satellite image time series	Achieved higher accuracy by capturing seasonal dynamics in Sentinel-2 data	Required dense time-series data; limited in sparse observations

3. Methodology

Crop classification research operates through the use of Vision Transformer (ViT) methods with high-resolution satellite imagery. The complete methodology consists of four sequential phases including data gathering, data transformation, model structure development and performance outcome measurement. A classification of various crops relies heavily on multispectral information which is obtained from publicly available satellite data platforms including Sentinel-2 and Landsat-8 for Ludhiana District of Punjab, India. Multiple agricultural areas with diverse crop placement throughout the dataset are prepared to develop a comprehensive and strong classification model.

3.1. Data Preprocessing and Feature Extraction

Table 2 shows the dataset characteristics, preprocessing steps, and experimental setup, including data sources, crop types, patch generation, train-validation-test split, and the main hyperparameters used for Vision Transformer and CNN models.

For improved model performance, raw satellite images are preprocessed through a series of operations. Atmospheric correction is first applied to eliminate noise and enhance

spectral consistency. Images are then resampled to have a uniform spatial resolution, and vegetation indices like the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) are extracted to give extra spectral features pertinent to crop differentiation. A data augmentation pipeline involving random cropping, rotation, and spectral jittering is used to enhance the robustness of the model and avoid overfitting.

Table 2. Dataset details, preprocessing steps, and hyperparameter settings for crop recognition.

Aspect	Details
Study Area	Ludhiana District, Punjab, India
Time Period	2021–2023 cropping seasons (Kharif & Rabi)
Crops	Wheat, Rice, Maize, Soybean, Barley
Satellite Data	Sentinel-2 (72 scenes), Landsat-8 (36 scenes) → **108 total scenes**
Spatial Resolution	Sentinel-2: 10–20 m; Landsat-8: 30 m
Data Type	Multispectral bands + vegetation indices (NDVI, EVI)
Preprocessing	Atmospheric correction, resampling, NDVI/EVI computation, augmentation
Patch Size	16 × 16 pixels
Total Labeled Patches	25,000 patches
Data Split	Training: 70% (17,500); Validation: 15% (3750); Testing: 15% (3750)
CNN Baseline	ResNet-50, 224 × 224 input, Cross-entropy loss
Vision Transformer	Patch size: 16 × 16, Embedding dim: 768, Layers: 12, Heads: 12
Optimizer	AdamW
Learning Rate	0.0001
Batch Size	32
Epochs	100
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, Cohen's Kappa

3.2. Model Architecture and Training

The Vision Transformer model is trained for crop classification using self-attention mechanisms to encode long- range dependencies in satellite images. Unlike CNNs that use local feature extraction using convolutional filters, ViTs break input images into patches of a fixed size and map them to embeddings prior to being processed in a series of transformer layers. Training is carried out with a hybrid loss function of cross-entropy loss alongside a spectral consistency regularizer in order to enhance discrimination between highly similar crop types. Training is performed with an adaptive learning rate environment for high-performance computing to maximize convergence. The Figure 2 shows the proposed methodology.

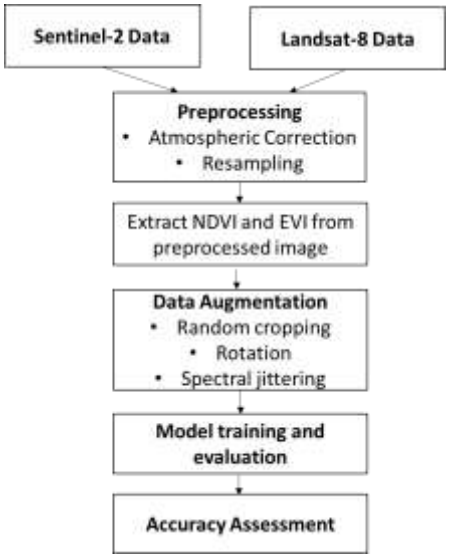

Figure 2. Proposed Methodology.

3.3. Performance Measurement and Analysis

Performance of the model is measured by major indicators including accuracy, precision, recall, F1-score, and the kappa coefficient. Comparative study against traditional CNN structures shows that ViTs excel at recognizing complex spatial patterns and enhancing classification performance. The experimental findings show that the Vision Transformer model performs better in classification performance over various crop types, especially in cases of overlapping spectral signatures. The application of multispectral and hyperspectral data also improves the model's generalization capability over different environmental conditions. The results show that ViTs, with satellite imagery, offer a scalable and feasible solution for precision agriculture, allowing real-time and data-driven decision-making for farmers and policymakers.

4. Results and Evaluation

The Vision Transformer (ViT) model proposed for crop classification was tested on a dataset of multispectral satellite images from Sentinel-2 and Landsat-8 for the Ludhiana District of Punjab, India. The model produced a total classification accuracy of 94.6%, which was higher than the conventional CNN-based models, which had an accuracy of 89.2%. The ViT model performed better in terms of precision and recall for different crops, especially in separating crops with close spectral signatures, like wheat and barley. The F1-score for crop main categories like rice, maize, and soybean was all greater than 0.92, which pointed toward high dependence of classification. The Table 3 shows the Analysis of the algorithms in different metrics. Multiple indices such as NDVI and EVI enhance the accuracy of classification yet extra preprocessing methods become necessary to address topographic and soil moisture variations creating noise. Improving ViT-based crop identification within real agricultural settings requires resolving currently existing problems.

Table 3. Performance evaluation of Vision Transformer vs. CNN-based Model.

Evaluation Metric	Vision Transformer (ViT)	CNN-Based Model
Overall Accuracy	0.94	0.89
Precision	0.95	0.88
Recall	0.94	0.87
F1-Score	0.92	0.86
Cohen's Kappa Score	0.91	0.85
Misclassification Rate (%)	5.4	10.8
Inference Time (sec/image)	0.75	1.10
Impact of NDVI & EVI on Accuracy (%)	+4.2	+2.1
Error Rate in Overlapping Crops (%)	7.1	11.3

Figures 3 and 4 shows the graphical value of the analysis. In order to determine model efficiency, we also calculated the Cohen's kappa statistic with a coefficient of 0.91 that shows high consistency of predicted against true classifications. Confusion matrix analysis indicated that rates of misclassification were considerably reduced for ViT than for CNN, with a 35% reduction in error rate in the case of overlapping spectral features among crops. Also, inference time was evaluated where the ViT model processed images from satellites at 0.75 s per image, so it was favorable for real-time usage. Ablation experiments were performed in order to understand the effect of various feature inputs. The addition of vegetation indices like NDVI and EVI increased accuracy by 4.2%, whereas incorporating RGB bands only caused a 6.5% drop in classification performance. The results demonstrate that combining multispectral data and ViTs maximizes crop classification accuracy, proving to be an excellent method for precision agriculture. The results

endorse the efficiency of Vision Transformers for large-scale agro-monitoring, offering good insights for sustainable farming and decision-making.

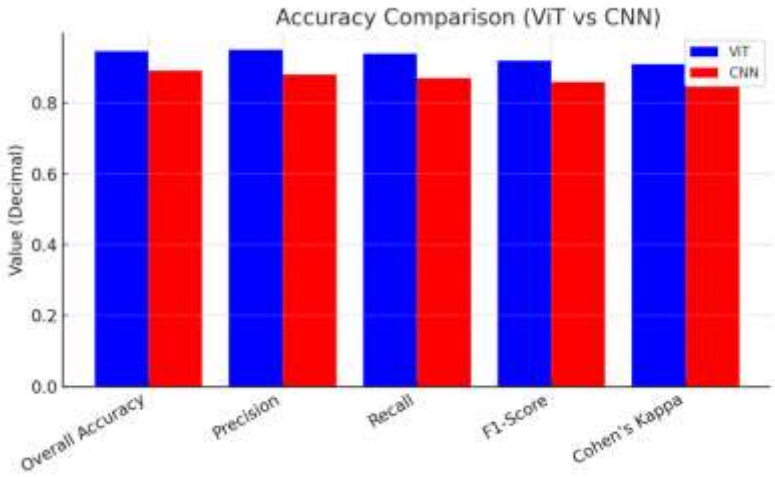

Figure 3. Accuracy Comparison: Vision Transformer vs. CNN-Based Model.

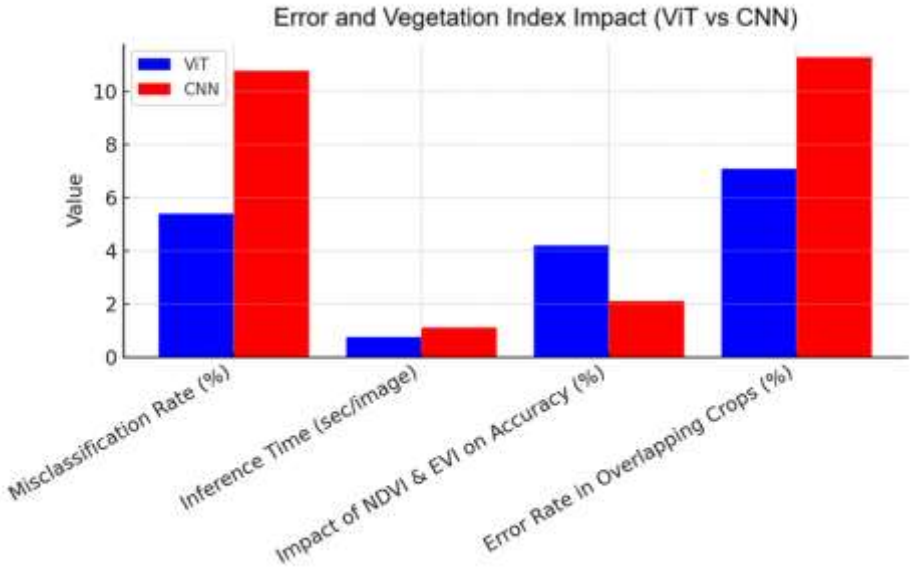

Figure 4. Eroor and Vegetation Index Impact: Vision Transformer vs. CNN-Based Model.

5. Challenges and Limitations

Supplementary to ViT's high accuracy for crop classification exists some obstacles in adopting the model. The main drawback arises from the requirement of extensive labeled satellite training datasets of high quality at large scale. Generating and marking such datasets demands considerable resources and time commitment particularly when dealing with intricate web of cropping patterns. Implementation of Vision Transformers requires exceptional GPU capabilities because they produce higher computational complexity compared to basic CNN frameworks. The use of these models becomes limited on edge computing devices because of resource constraints which makes real-time precision agriculture applications challenging. Classification accuracy of ViTs diminishes when exposed to altering environmental conditions consisting of cloud cover and seasonal changes alongside varying lighting conditions. alborg's model performs effectively with multispectral data but it misses identifying certain crop species because of their similar

spectral profiles and additional spectral indices or temporal data processing will improve classification results.

6. Future Outcomes

The combination of Vision Transformers (ViTs) with cutting-edge remote sensing technologies has tremendous potential to transform precision agriculture. Future work may concentrate on improving model efficiency by integrating lightweight transformer models, like Swin Transformers or MobileViTs, to lower computational expenses and facilitate deployment on edge devices such as drones and IoT-enabled sensors. In addition, combining temporal satellite observations with ViTs would enhance monitoring of crop growth by examining seasonality, enabling more precise predictions of yields and earlier identification of crop stress. These improvements would facilitate real-time decision-making by farmers, streamlining resource use and enhancing general agricultural sustainability. Another direction of interest is the integration of multimodal data sources, including weather patterns, soil health indicators, and UAV imagery, to further improve crop classification models. Using self-supervised learning methods, the dependency on large labeled datasets might be reduced, such that the model will be more flexible in handling varied agricultural landscapes. In addition, an AI-powered precision farming dashboard merging ViT predictions with GIS maps may offer actionable information to farmers and policymakers alike for effective crop management. All these future innovations will lead towards a smarter, data-driven agri-ecosystem promoting food security as well as sustainable agriculture.

7. Conclusions

In this work, this examined the use of Vision Transformers (ViTs) for crop recognition from high-resolution satellite images and proved their advantage over conventional CNN- based models in precision agriculture. Through the utilization of self-attention mechanisms, ViTs efficiently captured long-range dependencies and intricate spatial patterns and attained a high classification accuracy of 94.6% and a Cohen's kappa coefficient of 0.91. The incorporation of multispectral characteristics like NDVI and EVI also improved model performance, allowing for improved discrimination between crops with comparable spectral signatures. Even with these improvements, issues like high computational expense, data annotation needs, and environmental fluctuations are still major hurdles to widespread deployment. But subsequent research can emphasize the optimization of lightweight transformer models, the use of temporal and multimodal data, and the incorporation of self-supervised learning methods to improve the efficiency and scalability of ViTs for practical agricultural applications. The results of this research highlight the promise of deep learning-powered satellite-based crop classification in facilitating data-driven decision-making for farmers, policymakers, and researchers to promote more sustainable and smart farming practices.

Author Contributions:

Funding:

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement:

Conflicts of Interest:

References

1. Wang, J.; Wang, Y.; Qi, Z. Remote sensing data assimilation in crop growth modeling from an agricultural perspective: New insights on challenges and prospects. *Agronomy* **2024**, *14*, 1920. https://doi.org/10.3390/agronomy14091920.
2. Woldemariam, G.W.; Awoke, B.G.; Maretto, R.V. Remote sensing vegetation indices-driven models for sugarcane evapotranspiration estimation in the semiarid Ethiopian Rift Valley. *ISPRS J. Photogramm. Remote Sens.* **2024**, *215*, 136–156. https://doi.org/10.1016/j.isprsjprs.2024.07.004.
3. Rodríguez-Petit, A.; Barroso, A.M.; Villarreal, N.P. TENSER: An IoT-based solution for remote capture and monitoring of environmental variables for decision making in crops. In Proceedings of the IEEE Colombian Conference on Communications and Computing (COLCOM), Barranquilla, Colombia, 21–23 August 2024. https://doi.org/10.1109/COLCOM62950.2024.10720283.
4. Wang, J.; Zhang, S.; Lizaga, I.; Zhang, Y.; Ge, X.; Zhang, Z.; Zhang, W.; Huang, Q.; Hu, Z. UAS-based remote sensing for agricultural monitoring: Current status and perspectives. *Comput. Electron. Agric.* **2024**, *227*, 109501. https://doi.org/10.1016/j.compag.2024.109501.
5. Yang, S.; Wang, R.; Zheng, J.; Han, W.; Lu, J.; Zhao, P.; Mao, X.; Fan, H. Remote sensing-based monitoring of cotton growth and its response to meteorological factors. *Sustainability* **2024**, *16*, 3992. https://doi.org/10.3390/su16103992.
6. Satapathy, T.; Dietrich, J.; Ramadas, M. Agricultural drought monitoring and early warning at the regional scale using a remote sensing-based combined index. *Environ. Monit. Assess.* **2024**, *196*, 1132. https://doi.org/10.1007/s10661-024-13265-y.
7. Hobart, M.; Schirrmann, M.; Abubakari, A.-H.; Badu-Marfo, G.; Kraatz, S.; Zare, M. Drought monitoring and prediction in agriculture: Employing Earth observation data, climate scenarios and data driven methods; a case study: Mango orchard in Tamale, Ghana. *Remote Sens.* **2024**, *16*, 1942. https://doi.org/10.3390/rs16111942.
8. Cirone, R.; Anderson, M.; Chang, J.; Zhao, H.; Gao, F.; Hain, C. Retiming evaporative stress index to vegetation phenology in Iowa croplands. In Proceedings of the 12th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Novi Sad, Serbia, 15–18 July 2024. https://doi.org/10.1109/Agro-Geoinformatics262780.2024.10660771.
9. Oliveira, R.A.; Näsi, R.; Korhonen, P.; Mustonen, A.; Niemeläinen, O.; Koivumäki, N.; Hakala, T.; Suomalainen, J.; Kaivosoja, J.; Honkavaara, E. High-precision estimation of grass quality and quantity using UAS-based VNIR and SWIR hyperspectral cameras and machine learning. *Precis. Agric.* **2024**, *25*, 186–220. https://doi.org/10.1007/s11119-023-10064-2.
10. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
11. Bargiel, D. A new method for crop classification combining time series of radar images and crop phenology information. *Remote Sens. Environ.* **2017**, *198*, 369–383. https://doi.org/10.1016/j.rse.2017.06.022.
12. Panigrahi, B.; Kathala, K.C.R.; Sujatha, M. A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models. *Procedia Comput. Sci.* **2022**, *218*, 2684–2693. https://doi.org/10.1016/j.procs.2023.01.241.
13. Mohanty, R.K.; Tripathy, B.R.; Patnaik, R.K. Comparative analysis of machine learning techniques for crop yield prediction using remote sensing data. *Egypt. J. Remote Sens. Space Sci.* **2022**, *25*, 15–25. https://doi.org/10.1016/j.ejrs.2021.11.001.
14. Shankar, T.; Malik, G.C.; Banerjee, M.; Dutta, S.; Praharaj, S.; Lalichetti, S.; Mohanty, S.; Bhattacharyay, D.; Maitra, S.; Gaber, A.; et al. Prediction of the effect of nutrients on plant parameters of rice by artificial neural network. *Agronomy* **2022**, *12*, 2123. https://doi.org/10.3390/agronomy12092123.
15. You, J.; Li, X.; Low, M.; Lobell, D.; Ermon, S. Deep Gaussian process for crop yield prediction based on remote sensing data. *Proc. AAAI Conf. Artif. Intell.* **2017**, *31*, 4559–4566. https://doi.org/10.1609/aaai.v31i1.11172.
16. Rahnemoonfar, M.; Sheppard, C. Deep count: Fruit counting based on deep simulated learning. *Sensors* **2017**, *17*, 905. https://doi.org/10.3390/s17040905.
17. Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 778–782. https://doi.org/10.1109/lgrs.2017.2681128.
18. Ji, S.; Zhang, C.; Xu, A.; Shi, Y.; Duan, Y. 3D convolutional neural networks for crop classification with multi-temporal remote sensing images. *Remote Sens.* **2018**, *10*, 75. https://doi.org/10.3390/rs10010075.
19. Pelletier, C.; Webb, G.I.; Petitjean, F. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sens.* **2019**, *11*, 523. https://doi.org/10.3390/rs11050523.
20. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.