International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

A Spatio-Temporal Transformer Framework for High-Resolution Crop Yield Prediction Across Agro-Climatic Zones: Survey

A. Srilatha
Research Scholar, Department of CSE, Bharatiya Engineering Science and Technology Innovation University, Gownivaripalli , Gorantla in Andhra Pradesh

Dr. Ranga Swamy Sirisati
Associate Professor , Department of CSE , Vignan's Institute Of Management and Technology For Women, Ghatkesar, Medchal, Telangana

Abstract : Accurate crop yield prediction is a cornerstone of modern precision agriculture, enabling proactive decision-making for food security, supply-chain optimization, and climate-resilient farming. Traditional statistical models and conventional machine learning approaches are limited in their ability to model complex, nonlinear, and long-range dependencies inherent in agro-climatic systems. Recent advances in deep learning, particularly transformer-based architecture, have demonstrated remarkable performance in modeling spatio-temporal dependencies across diverse domains. This paper presents an extensive study on crop yield prediction using a spatio-temporal transformer framework capable of integrating multi-modal datasets such as satellite imagery, weather observations, soil characteristics, and historical yield records. A comprehensive literature survey is conducted to examine existing machine learning, deep learning, and attention-based models. Based on the identified research gaps, multiple transformer-driven methodologies are proposed with mathematical formulations. Comparative analysis using reported metrics from prior studies highlights the superiority of transformer-based models in terms of accuracy, scalability, and robustness across agro-climatic zones.

Keywords:

Crop Yield Prediction, Spatio-Temporal Transformers, Precision Agriculture, Deep Learning, Multi-Modal Data

1. INTRODUCTION

Agriculture remains one of the most critical sectors for global economic stability and human sustenance. Accurate prediction of crop yield plays a vital role in food security planning, market regulation, insurance assessment, and climate adaptation strategies. However, crop productivity is influenced by a complex interplay of spatial and temporal factors including soil conditions, weather patterns, vegetation dynamics, and agronomic practices. Traditional yield prediction approaches relied heavily on empirical statistical models and linear regression techniques, which are often insufficient under dynamic climatic conditions. Machine learning models such as Random Forests and Support Vector Machines improved prediction accuracy but still struggle with temporal dependencies and spatial heterogeneity. Deep learning models, particularly CNN-LSTM hybrids, partially address these challenges but are limited by fixed receptive fields and vanishing gradient issues.

Transformer architecture, originally proposed for natural language processing, employ self-attention mechanisms that allow efficient modeling of long-range dependencies.
Recent research indicates that spatio-temporal transformers can effectively capture complex interactions in agricultural datasets, making them highly suitable for high-resolution crop yield prediction across diverse agro-climatic zones.

2. DATASETS FOR CROP YIELD PREDICTION

Crop yield prediction relies on heterogeneous datasets collected from multiple sources and scales. This study considers widely used and publicly available datasets to ensure reproducibility and scalability.

1. **USDA NASS Dataset:** Provides county-level crop yield statistics for maize, wheat, and soybean across multiple decades.
2. **MODIS Satellite Data:** Offers vegetation indices such as NDVI and EVI with high temporal frequency, enabling crop health monitoring.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

3. **ERA5 Climate Reanalysis Dataset:** Supplies hourly and daily meteorological variables including temperature, precipitation, wind speed, and solar radiation.
4. **Soil-Grids Dataset:** Contains global soil property maps including organic carbon content, texture, moisture, and pH values.
5. **Indian Ministry of Agriculture Statistics:** Provides district-wise seasonal crop yield data useful for regional modeling.

The integration of these datasets enables comprehensive spatio-temporal representation of agricultural systems.

3. LITERATURE SURVEY

Numerous studies have explored crop yield prediction using data-driven approaches. Early studies employed statistical regression and time-series models such as ARIMA, which were limited by linear assumptions. Machine learning techniques including Random Forests, Gradient Boosting, and Support Vector Regression improved performance by capturing nonlinear relationships. Deep learning approaches marked a significant shift in yield prediction research. Convolutional Neural Networks were applied to satellite imagery for spatial feature extraction, while Recurrent Neural Networks and LSTMs modeled temporal dependencies. Hybrid CNN-LSTM models demonstrated improved accuracy but suffered from scalability and long-term dependency limitations. Recent works have incorporated attention mechanisms and transformer architectures. Spatio-temporal attention models have shown enhanced interpretability and performance. Transformers, with their global self-attention capability, have emerged as a promising solution for high-resolution, multi-modal crop yield forecasting.

4. RESEARCH GAPS

Despite significant advancements, several challenges remain unresolved:

1. Limited adoption of transformer architectures tailored for agricultural datasets.
2. Poor generalization of existing models across heterogeneous agro-climatic zones.
3. Inadequate fusion of multi-modal data sources.
4. Lack of explainability in deep learning-based yield prediction models.
5. High computational complexity and data scarcity in developing regions.

5. PROBLEM STATEMENT

Existing crop yield prediction models are unable to effectively capture complex spatio-temporal dependencies and multi-modal interactions present across diverse agro-climatic zones. This results in suboptimal prediction accuracy, limited scalability, and poor interpretability. There is a critical need for a robust spatio-temporal transformer-based framework capable of delivering high-resolution, accurate, and generalizable crop yield predictions.

6. PROPOSED METHODOLOGIES

6.1 Spatio-Temporal Transformer Architecture:

In crop yield prediction, agricultural data is naturally **spatial and temporal** — weather evolves over time and varies across regions. To capture both dimensions concurrently, we structure the input data as a three-dimensional tensor: data is collected over time and across different geographical locations, with multiple environmental features.

To represent this information in a structured form, the input data is organized as a three-dimensional matrix.

Input Representation

The input dataset is represented as:

$$\mathbf{X} \in \mathbb{R}^{T \times S \times F}$$

Where:

- **T (Temporal steps):** Number of time intervals such as weeks, months, or crop growth stages.
- **S (Spatial locations):** Number of geographical units such as fields, districts, or grid cells.
- **F (Feature dimensions):** Number of input variables such as temperature, rainfall, NDVI, soil moisture, and humidity.

In simple words: This representation allows the model to simultaneously learn:

- How crop yield changes over time
- How crop yield varies across locations

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

- How different environmental features influence yield

This spatio-temporal representation enables the model to learn patterns in crop growth over time and across different locations. Similar multi-dimensional formulations have been used effectively in recent deep learning crop yield research to handle complex dependencies from weather, satellite, and soil data simultaneously.

Self-attention: The transformer uses a self-attention mechanism to learn important relationships within the data. Self-attention helps the model understand which time periods and locations are most influential for predicting crop yield.

The core innovation in transformer models is the self-attention mechanism, which dynamically learns the importance of different data points:

The attention operation is mathematically expressed

$$\text{Attention}(Q, K, V) = \text{softmax}((QK^T) / \sqrt{d_k}) \times V$$

In this formula:

- Q (Query) represents the current temporal and spatial context asking *which parts of the input are most influential*.
- K (Key) encodes all input contexts,
- V (Value) carries the actual feature representations,
- $d_k$is a scaling factor that stabilizes training.

This mechanism enables the model to capture long-range dependencies, which are critical for crop models where events like early-season droughts influence final yield.

The attention mechanism:

1. Compares the current data with all other data points
2. Assigns higher importance to relevant time periods or regions
3. Aggregates useful information for accurate yield prediction

In agriculture: This allows the model to focus more on critical growth stages, extreme weather events, or important spatial regions.

6.2 Positional Encoding: Temporal order is preserved using positional embeddings:

Unlike traditional neural networks, transformers do not automatically understand the order of time. However, in agriculture, seasonal and crop growth stages are extremely important.

To solve this, positional encoding is added:

$$Z_t = X_t + P_t$$

Explanation in Simple Words

- **X_t:** Original input features at time step *t*
- **P_t:** Positional information representing the time order (e.g., sowing stage, flowering stage, harvesting stage)

Why this is important: It helps the model distinguish between:

- Early-season rainfall
- Mid-season drought
- Late-season temperature stress

Thus, the model learns seasonal patterns effectively.

6.3 Multi-Modal Feature:

Crop yield prediction depends on multiple data sources, including:

- Satellite imagery (crop health)
- Weather data (rainfall, temperature)
- Soil properties (moisture, nutrients)

Instead of simply combining these features, a cross-attention mechanism is used.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

This is represented as:

$$\mathbf{Z = Attention(Q_{satellite}, K_{weather}, V_{weather})}$$

Where;

- Satellite data acts as the primary reference
- Weather and soil data provide supporting context
- The model learns how weather and soil conditions influence crop health seen in satellite images
- $Q_{satelite}$ is the satellite feature query,
- $K_{weather}, V_{weather}$ are key and value pairs from weather/soil modalities.
- Cross-attention enables the model to learn how changes in weather influence satellite-observed crop health patterns. Recent multimodal agricultural models demonstrate that cross-modal attention significantly improves prediction robustness and helps attribute importance to different data sources.

Practical Meaning: For example:

- Low NDVI combined with low rainfall → stress condition
- High NDVI with adequate rainfall → healthy growth

This intelligent fusion improves prediction accuracy significantly.

6.4 Yield Prediction Layer:

After learning complex spatio-temporal patterns, the model converts the learned representation into a final yield value.

This is expressed as:

$$\mathbf{\hat{y} = WZ + b}$$

Explanation in Simple Words

- **Z:** Learned features after attention and fusion
- **W and b:** Model parameters
- **ŷ:** Predicted crop yield

In simple terms: This layer translates learned agricultural patterns into an actual yield estimate such as tons per hectare.

7. COMPARATIVE ANALYSIS

Reported results from the literature indicate clear performance trends:

1. Random Forest models report RMSE values between 0.45–0.60.
2. CNN-LSTM models achieve RMSE between 0.30–0.45.
3. Attention-based RNNs reduce RMSE to approximately 0.25–0.35.
4. Transformer-based models demonstrate superior performance with RMSE values as low as 0.18–0.28.

These results confirm the effectiveness of transformer architectures in modeling complex agro-climatic relationships.

Recent crop yield prediction research highlights the evolution from traditional machine learning to deep learning and transformer methods:

Model / Approach	Performance	Key Insight
CNN-LSTM with Attention (2024)	RMSE ~0.017, $R^2 = 0.967$	Hybrid deep model performs well on wheat and rice datasets.
Multi-Modal Deep Ensemble (2025)	MAE 341 kg/Ha	RicEns-Net integrates SAR, optical and weather data via ensemble learning.
Transformer Models (2024)	Improved R^2 and predictive reliability	Transformer-based architectures outperform CNN/LSTM baselines.
Crossformer (2025)	R^2 ~0.9863	Cross-window attention enables superior spatio-temporal modeling.
Explainable Transformer (2025)	R^2 higher than CNN & RNN	Intrinsic explainability via attention supports interpretation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

These studies demonstrate that integrating advanced deep learning, ensemble learning, and transformer mechanisms leads to **significant performance improvements** compared to classical ML or hybrid CNN–LSTM methods.

8. CONCLUSION AND FUTURE WORK

This paper presented an in-depth exploration of spatio-temporal transformer frameworks for crop yield prediction. By analyzing datasets, reviewing literature, identifying gaps, and proposing advanced methodologies, the study establishes a strong foundation for future research.
Future work will focus on explainable AI integration, federated learning for privacy-preserving agriculture, and real-time deployment using IoT-enabled platforms.

REFERENCES

[1] You, J., et al., Deep Gaussian Process for Crop Yield Prediction, AAAI, 2017.
[2] Khaki, S., Wang, L., Crop Yield Prediction Using Deep Neural Networks, Frontiers in Plant Science, 2019.
[3] Liakos, K.G., et al., Machine Learning in Agriculture, Sensors, 2018.
[4] Garnot, V.S.F., et al., Satellite Image Time Series with Transformers, NeurIPS, 2020.
[5] Cai, Y., et al., Spatio-Temporal Attention for Crop Yield Prediction, IEEE TGRS, 2022.
[6] Shahhosseini, M., et al., Yield Prediction Using Machine Learning, PLOS ONE, 2021.
[7] S. Y. Pravesh et al., "Predictive Modeling of Crop Yield Using Deep Learning Based Transformer With Climate Change Effects," *International Research Journal of Multidisciplinary Technovation*, 2024.
[8] Priyadarshini, P. M. Talwar, U. P. Rathod, S. Kulkarni, and B. R. Mulge, "Deep-Learning Based Crop Yield Prediction Model For Optimizing Agricultural Productivity And Food Security," *J. Scientific Research and Technology*, Jun. 2025.
[9] D. Yewle, L. Mirzayeva, and O. Karakuş, "Multi-modal Data Fusion and Deep Ensemble Learning for Accurate Crop Yield Prediction," *arXiv*, Feb. 2025.
[10] H. Najjar, D. Pathak, M. Nuske, and A. Dengel, "Intrinsic Explainability of Multimodal Learning for Crop Yield Prediction," *arXiv*, Aug. 2025.
[11] S. Dangi et al., "A multi-temporal multi-spectral attention-augmented deep CNN for crop yield prediction," *arXiv*, Sep. 2025.
[12] H. Kamangir et al., "CMAViT: Integrating Climate, Management, and Remote Sensing Data for Crop Yield Estimation with Multimodel Vision Transformers," *arXiv*, Nov. 2024.
[13] R. Sharma, J. Kaur, G. Feng, et al., "Maize and soybean yield prediction using machine learning methods: a systematic literature review," *Discov. Agric.* 3, 64 (2025).
[14] Y. Wang et al., "Progress in Research on Deep Learning-Based Crop Yield Prediction," *Agronomy*, 14(10), Oct. 2024.