A Multimodal Deep Learning Framework for Crop Yield Prediction from Satellite and Weather Data

Sofia R. Albrecht^{1,2,#}, Daniel O. Katsaros^{1,#}, Amara J. Okonkwo^{1,2}, Henrik P. Solberg¹, Leila M. Haddad^{1,2,*}

¹Department of Agricultural Data Science, Eastvale University of Applied Sciences, Korsholm, Finland; ²Centre for Precision Agriculture Analytics, Eastvale University of Applied Sciences, Korsholm, Finland

Correspondence to: Dr. Leila M. Haddad, Department of Agricultural Data Science, Eastvale University of Applied Sciences, Korsholm, Finland; Email: l.haddad@eastvale.example.edu

ABSTRACT

Reliable crop yield prediction requires integrating static field characteristics with time-varying satellite and weather observations, yet most operational models treat these sources separately. This study proposes a multimodal deep learning architecture, combining a temporal convolutional network, a bidirectional gated recurrent unit (GRU) branch, and a cross-attention fusion module, for yield class prediction across heterogeneous inputs. The proposed model was evaluated against three established baselines (logistic regression, random forest, and CatBoost) on a combined field-level and seasonal time-series benchmark of 54,200 records. The proposed model achieved 93.2% accuracy and an F1-score of 0.925, outperforming the strongest baseline (CatBoost) by 5.2 percentage points, at a moderate increase in inference time (12.6 ms versus 8.1 ms). Ablation analysis showed that removing the bidirectional GRU branch reduced accuracy by 5.5 points and removing the cross-attention module reduced accuracy by 3.1 points, indicating that both temporal modeling and cross-modal fusion contribute meaningfully to overall performance. These results suggest that multimodal architectures offer a practical and accurate approach for yield prediction tasks involving mixed static and sequential data.

Keywords: Crop yield prediction · Precision agriculture · Multimodal learning · Cross-attention · Remote sensing · Deep learning

INTRODUCTION

Climate variability and growing food demand have increased the need for accurate, field-level yield forecasts that can support planning, insurance, and supply-chain decisions. Operational forecasting pipelines typically combine two types of information: static field characteristics such as soil properties and management practices, and dynamic signals such as satellite vegetation indices and daily weather records. Conventional approaches handle these inputs separately, which can limit their ability to capture interactions between field conditions and seasonal crop development [2;5].

Tree-based ensemble methods, including random forest and gradient boosting, remain widely used for tabular agricultural data because of their robustness, interpretability, and low computational cost [3;5]. However, these methods rely on hand-crafted seasonal summary statistics to represent time-dependent behavior, which may discard fine-grained temporal patterns, such as the timing of a dry spell during flowering, that strongly influence yield. Deep learning architectures combining convolutional and recurrent components have shown promise for remote-sensing time series, but their integration with static field data through explicit fusion mechanisms has been comparatively underexplored [6;8].

This study proposes a multimodal architecture designed to jointly process static field features and satellite and weather sequences for yield class prediction. The purpose of this study was to evaluate the predictive accuracy of the proposed model relative to established baselines, and to quantify the contribution of its individual components through ablation analysis.

© 2026 The Information Science Journal

METHODS

Dataset

A combined benchmark dataset of 54,200 field-season records was assembled from field-level survey data and associated remote-sensing and meteorological time series drawn from a publicly available agricultural monitoring corpus. Each record included 30 static features (soil composition, elevation, previous crop, and management variables) and a corresponding 24-step sequence of 9 channels (vegetation indices and weather variables) sampled at roughly ten-day intervals across the growing season. The prediction target was a three-level yield class (low, medium, high) defined relative to regional averages. The dataset was partitioned into training (70%), validation (15%), and test (15%) subsets using stratified sampling to preserve class balance across yield categories.

Model architecture

The proposed model consists of three components. A temporal convolutional network with dilated convolutions first extracts local seasonal patterns from the sensor sequence. The resulting feature maps are passed to a two-layer bidirectional GRU to capture dependencies across the full growing season. A cross-attention module then uses an embedding of the static field features as the query and the GRU outputs as keys and values, allowing the model to weight the time steps most relevant to each field's conditions. The attended representation is concatenated with the static features and passed to a fully connected classification head.

Baseline models

Three baseline models were trained for comparison: logistic regression, random forest (400 estimators), and CatBoost (600 iterations, depth 6). All baselines were trained on the static field features together with engineered seasonal summary statistics of each sequence channel (mean, maximum, cumulative sum, and peak timing), consistent with standard practice for these model classes. Hyperparameters were tuned using grid search on the validation set.

Training and evaluation

The proposed model was trained using the Adam optimizer with an initial learning rate of 8e-4, a batch size of 128, and early stopping based on validation loss over a maximum of 100 epochs. Performance was evaluated on the held-out test set using accuracy, F1-score, and mean inference time per record. An ablation study was conducted by systematically removing the bidirectional GRU branch, the cross-attention module, and the temporal augmentation procedure (random time-step masking and jitter) applied during training.

RESULTS

Predictive performance

The proposed model achieved the highest performance among all evaluated approaches, with 93.2% accuracy and an F1-score of 0.925 on the held-out test set (Table 1). This represented an improvement of 5.2 percentage points in accuracy over the strongest baseline, CatBoost (88.0%), and a larger margin over random forest (84.3%) and logistic regression (74.9%). The improvement was accompanied by an increase in mean inference time, from 8.1 ms for CatBoost to 12.6 ms for the proposed model.

Model	Input	Accuracy (%)	F1-Score	Inference Time (ms)
Logistic Regression	Static + seasonal statistics	74.9	0.712	1.8

© 2026 The Information Science Journal

Model	Input	Accuracy (%)	F1-Score	Inference Time (ms)
(baseline)				
Random Forest	Static + seasonal statistics	84.3	0.826	5.2
CatBoost	Static + seasonal statistics	88.0	0.866	8.1
TCN-BiGRU-CrossAttention (proposed)	Static + raw time series	93.2	0.925	12.6

Table 1: Predictive performance of the proposed model compared with baseline models on the held-out test set.

Ablation analysis

Ablation experiments indicated that all three components contributed to overall performance (Table 2). Removal of the bidirectional GRU branch, leaving only the temporal convolutional pathway, produced the largest decrease, reducing accuracy by 5.5 percentage points to 87.7%. Removal of the temporal augmentation reduced accuracy by 3.6 points to 89.6%, and removal of the cross-attention module, replaced by simple concatenation of pooled features, reduced accuracy by 3.1 points to 90.1%. These results suggest that season-long temporal modeling contributed most substantially to predictive performance, while cross-modal attention and augmentation provided smaller but consistent additional gains.

Ablation Configuration	Accuracy (%)	Δ from Full Model
Full proposed model	93.2	—
Without cross-attention module	90.1	-3.1
Without BiGRU branch (TCN only)	87.7	-5.5
Without temporal augmentation	89.6	-3.6

Table 2: Ablation study isolating the contribution of individual components to model accuracy.

DISCUSSION

The results indicate that a multimodal architecture combining temporal convolution, bidirectional recurrent modeling, and cross-attention fusion can outperform established ensemble baselines for yield class prediction. The magnitude of improvement over CatBoost, the strongest baseline evaluated, suggests that the full seasonal trajectory contains predictive signal that is not entirely preserved by hand-crafted summary statistics.

The ablation analysis further clarifies the contribution of each component. The large decrease observed when removing the bidirectional GRU branch indicates that season-long dependencies are a primary driver of the model's advantage, while the cross-attention module appears to provide a secondary refinement by conditioning temporal weighting on field characteristics such as soil type. The contribution of augmentation, though smaller, was consistent with its role in improving robustness to cloud-related gaps in satellite coverage.

The increase in inference time associated with the proposed model, while modest in absolute terms, is unlikely to be a limiting factor for seasonal forecasting but may matter for large-scale regional deployments processing millions of fields. This study has several limitations, including

© 2026 The Information Science Journal

reliance on a single benchmark dataset covering a limited set of crops and a fixed train-validation-test split; future work should evaluate the architecture across additional regions and crop types and assess its robustness under extreme weather years not represented in the training data.

CONCLUSION

This study presented a multimodal architecture for crop yield prediction involving combined static field data and satellite and weather time series, and demonstrated that it outperforms conventional ensemble baselines by a meaningful margin. Ablation analysis confirmed that both temporal modeling and cross-modal attention contribute to the observed performance gains. These findings support the use of multimodal deep learning for precision agriculture, while highlighting the trade-off between predictive accuracy and computational cost that should inform deployment decisions.

DECLARATIONS

Funding: This study was supported by the Eastvale University of Applied Sciences Precision Agriculture Research Fund [EUAS-PARF-2026-009] and the Korsholm Regional Food Systems Grant [KRFS-3342].

Conflicts of Interest: The authors declare no conflicts of interest.

Data Availability Statement: The benchmark dataset and processed model outputs supporting the findings of this study are available from the corresponding author upon reasonable request.

Author Contributions: Conceptualization, Haddad L.M. and Albrecht S.R.; methodology, Katsaros D.O.; software, Okonkwo A.J.; formal analysis, Solberg H.P.; investigation, Albrecht S.R.; writing—original draft, Albrecht S.R.; writing—review and editing, Haddad L.M. and Katsaros D.O.; supervision, Haddad L.M. All authors have read and agreed to the published version of the manuscript.

REFERENCES

1. Albrecht S.R., Katsaros D.O. Deep learning for field-scale yield forecasting. J Agric Inform. 2025;12(1):9-27.
2. Okonkwo A.J., Solberg H.P. Tabular models in agronomy: accuracy and limits. Ann Precis Farming. 2024;6(2):88-104.
3. Haddad L.M., Albrecht S.R. Seasonal summary statistics versus raw satellite series for yield modeling. Remote Sens Crop Sci. 2025;8(3):140-158.
4. Katsaros D.O., Okonkwo A.J. Fusing static and dynamic covariates in agricultural prediction. Comput Agric Syst. 2024;14(4):231-249.
5. Solberg H.P., Haddad L.M. Boosted trees in operational crop forecasting: a practitioner survey. Data Agric Rev. 2023;4(1):45-62.
6. Albrecht S.R., Solberg H.P. Recurrent and convolutional networks for vegetation index time series. Neural Earth Obs. 2025;9(2):66-85.

© 2026 The Information Science Journal