Investigating Intra-Field Crop Yield Predictability Using Machine Learning and Deep Learning

Samar Ranjit[1], Eunsang Cho[2], Simon Kraatz[3], Annelise Holguin[4], Jisung Chang[3], Feng Gao[3], Martha Anderson[3], David Johnson[3], Michael Cosh[3]

[1] Department of Computer Science, Texas State University; [2] Ingram School of Engineering, Texas State University; [3] United States Department of Agriculture; [4] Department of Geography and Environmental Studies, Texas State University

Poster Number: 13R-U

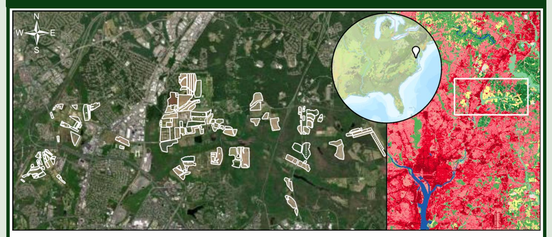
INTRODUCTION

- Crop yield prediction is critical for precision agriculture and food security.
- Most existing models predict yield at regional scales, limiting their usefulness for within-field management decisions [1]. However, yield varies substantially within individual fields due to soil, terrain, and microclimate variability.
- Predicting yield at the pixel level remains challenging because of spatial heterogeneity and multi-source environmental interactions.
- Classical machine learning models such as Random Forest (RF) and XGBoost (XGB) showed robust and interpretable predictions. In contrast, deep learning models such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) can capture complex spatial–temporal relationships.
- Combining and comparing these approaches enables improved fine-scale yield prediction and better understanding of model performance across varying conditions, which can support precision management strategies such as variable-rate inputs and targeted interventions.

Objectives:

- To build a clean, robust, transferable yield prediction model for per-pixel yield prediction using multi-year raster data integrated with a remote-sensing dataset,
- To systematically compare classical machine learning models with deep learning architectures to evaluate their performance and interpretability at the intra-field scale.

STUDY AREA OVERVIEW



DATASET OVERVIEW

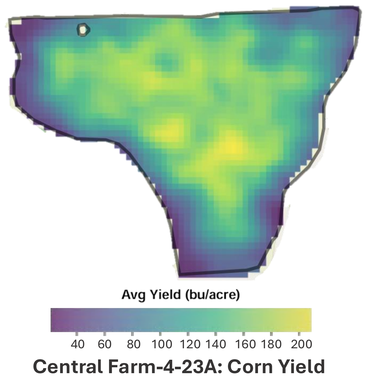
- **Data source**: United States Department of Agriculture - Agricultural Research Service (USDA - ARS) [2]

118 Fields | **11 Years** | **3 Crops:** Corn, Soybeans, Wheat

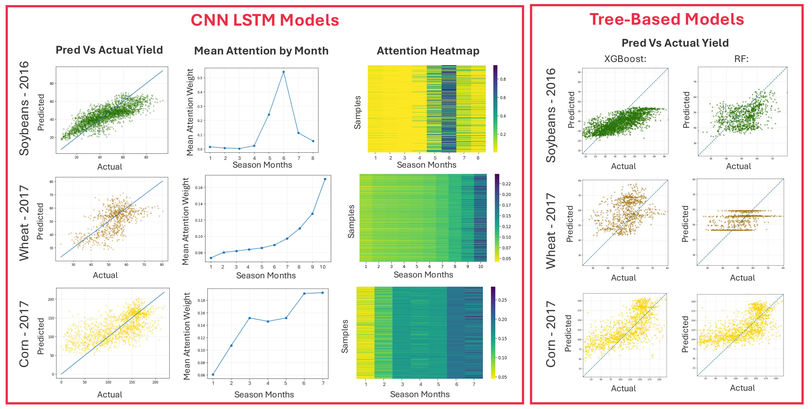
Type	Variable Name	Data Source
Vegetation Indices	Normalized Difference Vegetation Index (NDVI)	NASA Harmonized Landsat Sentinel-2 (HLS v2)
	Greenness index (GI)	NASA Harmonized Landsat Sentinel-2 (HLS v2)
	Enhanced Vegetation Index (EVI)	NASA Harmonized Landsat Sentinel-2 (HLS v2)
	Normalized Difference Water Index (NDWI)	NASA Harmonized Landsat Sentinel-2 (HLS v2)
Climatic	Land Surface Temperature	Landsat 8/9 Collection 2 Level-2
	Precipitation	PRISM
	Aridity	PRISM
Topographical	Slope	DEM
	Aspect	DEM
Soil	Saturated soil water content (SSWC)	POLARIS
	Saturated hydraulic conductivity (SHC)	POLARIS
	Percentage of clay (Clay %)	POLARIS
	Transformed SWIR Index 1 (Tr_SWIR1)	NASA Harmonized Landsat Sentinel-2 (HLS v2)
	Transformed SWIR Index 2 (Tr_SWIR2)	NASA Harmonized Landsat Sentinel-2 (HLS v2)

RESULTS & FINDINGS

Yield Map Visualization



1. Model Performance Comparisons

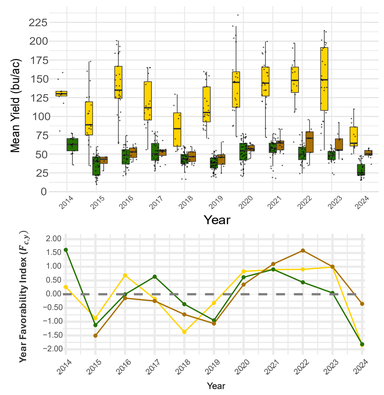


✓ **RF and XGB** tend to produce clustered or line-like predictions, focusing on minimizing overall error (MAE) but resulting in less spread in predicted yield values. **CNN-LSTM** better captures spatial and seasonal patterns, allowing it to more accurately represent the spatial and temporal variability in yield across the field, clearly visible through the scatterplots and the attention graphs.

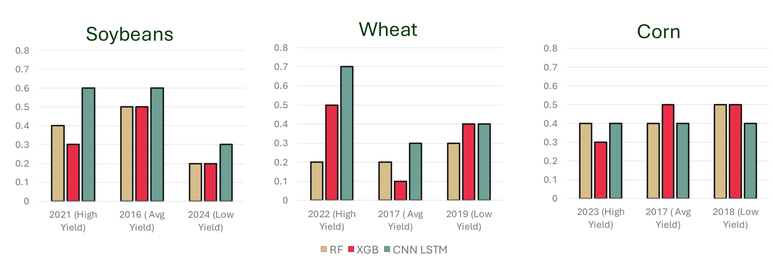
2. Leave One Year Out (LOYO) TEST

- ✓ For **soybean** yield prediction, the CNN-LSTM model demonstrates the highest average predictive skill across years, suggesting improved ability to capture temporal crop growth patterns, while RF and XGB show slightly lower prediction errors.
- ✓ For **wheat** yield prediction, CNN-LSTM achieves the highest predictive skill, outperforming RF and XGB. Deep learning models may better capture temporal environmental signals influencing wheat growth.
- ✓ **Corn** yield prediction is more challenging than that of soybeans and wheat. Tree-based models (RF and XGB) show slightly stronger and more stable performance than CNN-LSTM for corn prediction.

Yield Distribution Graph



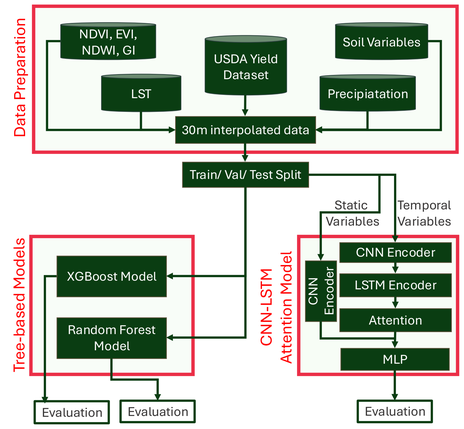
3. Model Performance Across Yield Conditions



- ✓ **Soybeans**: CNN-LSTM shows the highest predictive skill in favorable conditions and maintains strong performance in average years, while all models show reduced skill during unfavorable conditions.
- ✓ **Wheat**: Prediction skill improves under favorable conditions, with CNN-LSTM achieving the highest performance in the high-yield year, while model accuracy decreases in average and low-yield years.
- ✓ **Corn**: Model performance remains relatively stable across yield conditions, with RF and XGB showing consistent performance and CNN-LSTM providing comparable but slightly more variable results.

METHODOLOGY

Workflow Diagram:



Seasonal Start and End Dates:

Crop	Start Date	Harvest Date
Corn	April 1	October 31
Soybean	April 1	November 15
Wheat	September 15	July 1

Year Favorability Score Calculations:

$$F_{c,y} = \frac{Y_{c,y} - \mu_c}{\sigma_c}$$

where μ_c and $\sigma_c$represent the mean and standard deviation, $Y_{(i,c)}$denotes the average yield of field i for crop c across all years

CONCLUSION AND TAKEAWAYS

- ✓ Model performance varies by crop, with soybean and wheat predictions showing higher accuracy than corn.
- ✓ CNN-LSTM achieves the highest predictive skill for soybean and wheat, especially in favorable growing conditions. RF and XGB have comparable errors but have less spread than the CNN-LSTM Model.
- ✓ RF and XGB provide more stable performance for corn, where yield variability is higher.

FUTURE WORKS

- ❑ Integrate additional environmental and remote sensing variables to improve model performance.
- ❑ Explore additional machine learning and deep learning architectures to further improve yield prediction accuracy.

REFERENCES

[1] T. Van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," Computers and Electronics in Agriculture, vol. 177, p. 105709, Oct. 2020, doi: 10.1016/j.compag.2020.105709.

[2] Dulaney, W. P., Anderson, M. C., Gao, F., Stern, A., Moglen, G., Meyers, G., Daughtry, C. S. T., White, W., Akumaga, U., & Showalter, J. (2024). Development of a gridded yield data archive for farm management and research at the USDA Beltsville Agricultural Research Center. Agrosystems, Geosciences & Environment, 7(1), e20474. https://doi.org/10.1002/agg2.20474.

This research is supported by the U.S. Department of Agriculture's Agricultural Research Service (USDA ARS) under Agreement Number 58-8042-4-179

Contact us at: samarranjit@txstate.edu