Abstract

Advancements in agricultural data analytics, coupled with the increasing availability of multisource data, have enabled the development of intelligent systems capable of supporting sustainable and data-driven agricultural decision-making. However, existing predictive frameworks struggle with high-dimensional, multimodal, and incomplete datasets, limiting their applicability across diverse agro-climatic conditions. This research addresses these challenges by proposing robust, interpretable, and scalable deep learning methodologies for vegetation health monitoring, crop yield prediction, and growth stage detection. The thesis makes three key contributions: the development of a comprehensive NDVI forecasting pipeline, the design of multimodal and explainable yield prediction architectures, and the creation of a co-attention–based fusion framework for wheat growth stage recognition supported by a new multimodal dataset.

A comprehensive NDVI forecasting framework is developed using temporal smoothing, spatio-temporal modeling, and hybrid attention mechanisms. The Smoothed-LSTM model—integrated with Blackman FIR filtering—demonstrates substantial improvement in capturing nonlinear temporal patterns under noisy and incomplete satellite observations. Additional models, including a spatio-temporal LSTM for location-specific NDVI reconstruction and hybrid architectures combining Time2Vec embedding with GRU and attention layers, achieve significantly enhanced predictive performance. Ensemble strategies further reduce forecasting errors, making the approach robust across heterogeneous agricultural landscapes.

A multimodal crop yield prediction pipeline is introduced that integrates RGB, IR1, IR2, multispectral imagery, and environmental variables using a Meta-Transformer–based encoder followed by dimensionality reduction and Temporal Graph Neural Networks. This method effectively captures spectral, spatial, and temporal characteristics of crop growth, achieving state-of-the-art yield classification accuracy. SHAP-based explainability is incorporated to quantify feature importance, enabling interpretable dimensionality reduction and identification of key agronomic and meteorological factors influencing rice yield. A transfer learning–based model, YieldFusion, combining Time2Vec, TCN, Bi-LSTM, and multi-head attention, further enhances yield estimation accuracy in data-scarce regions by leveraging knowledge from the M4 dataset.

A Cross-Modal Co-Attention (CMCA) architecture is proposed to fuse Vision Transformer–based image representations with numerical features for accurate wheat growth stage recognition. The model significantly outperforms unimodal and concatenationbased baselines, demonstrating the importance of multimodal fusion for phenological understanding. To support this evaluation, a multimodal wheat dataset is developed consisting of RGB images, agronomic traits, and associated field metadata.

In summary, this research delivers end-to-end solutions for vegetation monitoring, yield forecasting, and growth stage detection using interpretable and generalizable deep learning methods. The findings support scalable precision agriculture by offering actionable insights for farmers, policymakers, and agronomists, ultimately contributing to improved productivity, resource efficiency, and food security.

Keywords: Precision Agriculture; Vegetation Health Monitoring; NDVI Forecasting; Multimodal Deep Learning; Crop Yield Prediction; Explainable AI (XAI); SHAP; Meta-Transformer; Temporal Graph Neural Networks; Transfer Learning; Growth Stage Detection; Cross-Modal Co-Attention; Vision Transformer; Multisource Remote Sensing; Data-Scarce Regions