Multimodal Transformer for Crop Monitoring in Portugal

Antonio Henrique Xavier da Silva

September 18, 2025

1 Study Problem

Portuguese agriculture faces growing challenges related to Mediterranean climatic variability, economic pressures from the Common Agricultural Policy (CAP), and the urgent need to increase productive efficiency in the context of climate change. Traditional crop monitoring, based on manual inspections and reactive strategies, proves inadequate to meet the demands of modern precision agriculture.

1.1 Context and Motivation

The Portuguese agricultural sector represents approximately 1.9% of national GDP, constituting a fundamental pillar of the national economy. However, Portuguese agricultural productivity is below the European average in several main crops, particularly in cereals (wheat, corn) and permanent crops (vineyards, olive groves), which represent more than 60% of the useful agricultural area.

The main challenges identified include:

Climatic Variability: The Mediterranean climate is characterized by significant inter-annual irregularity, with prolonged drought periods alternating with intense precipitation events. This variability makes traditional agronomic management difficult and increases the risks of water and thermal stress in crops.

Late Stress Detection: Conventional monitoring methods only allow reactive detection of crop problems, when damage is already visually evident and often irreversible. The ability for early stress detection is fundamental for effective interventions.

Spatial Heterogeneity: The Portuguese agricultural landscape presents high fragmentation and heterogeneity, with parcels of variable dimensions (average of 13.7 ha) and diverse edapho-climatic conditions, even at local scales. This heterogeneity requires spatially explicit monitoring approaches.

Multimodal Data Integration: Information relevant to agricultural management comes from multiple heterogeneous sources (satellite, meteorology, agronomic practices), whose effective integration remains a significant technical challenge.

1.2 Research Objectives

This research proposes the development of a transformer-based framework to specifically address these challenges through:

1. **Proactive Prediction**: Development of vegetation health indicator prediction capabilities up to 30 days in advance, allowing preventive interventions in optimal time windows.
2. **Multimodal Integration**: Adaptive fusion of Sentinel-2 data (5-day temporal resolution), meteorological data from IPMA (daily resolution), and phenological information (10-day resolution), exploring synergies between modalities.
3. **Regional Specialization**: Specific adaptation to Portuguese Mediterranean conditions, considering phenological patterns, agronomic practices, and regional climatic constraints.
4. **Agronomic Interpretability**: Development of explainability mechanisms that allow the translation of predictions into practical and scientifically grounded agronomic recommendations.

2 Study Area

2.1 Geographic Characterization

The study region focuses on Ribatejo, specifically the municipalities of Santarém, Cartaxo, and Almeirim, covering an intensive agricultural area in central Portugal. This region was selected based on the following criteria:

Data Availability: The IPMA meteorological network coverage is dense in the region and SIP data quality is high due to the region's importance in the CAP.

Edaphoclimatic Conditions: Predominantly alluvial soils with good water retention capacity, altitude between 10-150m, and Mediterranean climate with moderate Atlantic influence.

2.2 Regional Specificities

Phenological Patterns: The bimodal regime of crops (winter vs. spring cereals) creates complex temporal dynamics that require specific modeling. Corn shows peak development between July-August (DOY 200-230), while wheat reaches maximum vigor in April-May (DOY 100-130).

Water Regimes: The region presents structural summer water deficit (June-September), with annual potential evapotranspiration of 1,100mm vs. precipitation of 650mm. This condition makes water stress the main productivity limiting factor.

3 Model Architecture

3.1 General Concept

The proposed architecture implements a **Multimodal Spatio-Temporal Transformer (MSTT)** specifically designed to capture the complexities of Mediterranean agricultural systems. The model integrates three data sources through specialized attention mechanisms.

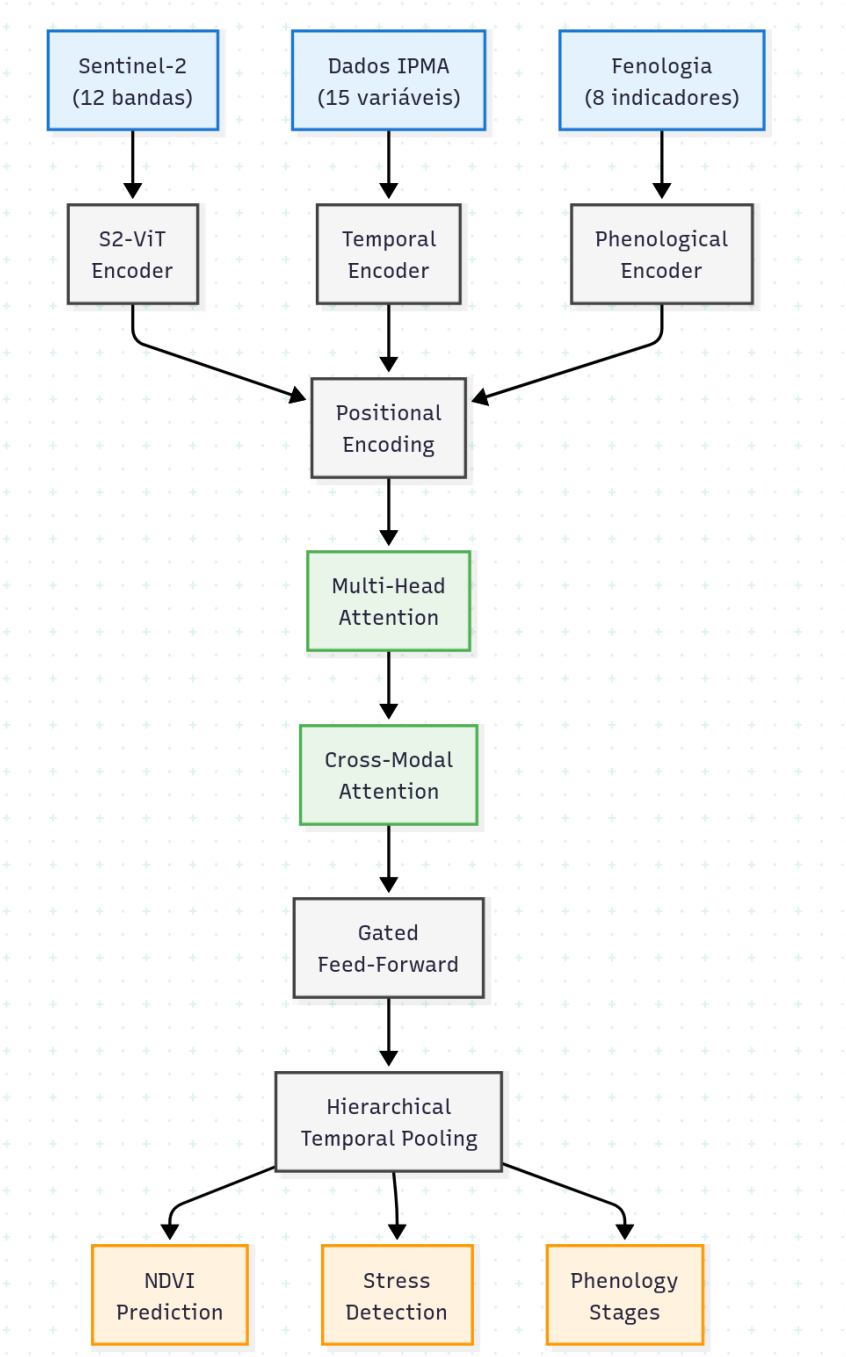

Figure 1: Multimodal Spatio-Temporal Transformer architecture for agricultural monitoring

3.2 Specialized Components

Sentinel-2 Vision Transformer (S2-ViT): Specific adaptation of the ViT architecture for multispectral data, incorporating spectral attention between the 12 Sentinel-2 bands. Patch embedding processes 16×16 pixel sub-regions, preserving local spatial information while allowing modeling of long-range dependencies.

Meteorological Temporal Encoder: Specialized encoder for meteorological time series with explicit seasonal decomposition. Processes 15 agrometeorological variables (temperature, precipitation, evapotranspiration, growing degree days) with cross-attention between variables to capture complex climatic interactions.

Phenological Feature Encoder: Specific component for phenological patterns derived from Copernicus VPP products, with specialized embeddings per crop and attention focused on critical development stages (emergence, flowering, maturation).

Cross-Modal Attention: Attention mechanism that allows each modality to *query* relevant information from other modalities. For example, spectral data can automatically focus on critical meteorological periods, while phenological information can identify the most informative spectral bands for each stage.

Hierarchical Temporal Pooling: Adaptive temporal aggregation that combines information at multiple temporal scales (daily, weekly, monthly) through learned attention weights, allowing capture of both rapid dynamics and seasonal trends.

4 Modeling Decisions

4.1 Rationale for Architectural Choices

Transformers vs. Recurrent Networks: The choice of transformer architecture is based on demonstrated superiority in capturing long-range temporal dependencies without the gradient vanishing problems of RNNs. For agricultural data, this is crucial for modeling carry-over effects between seasons and multi-annual phenological patterns.

Multimodality: Native integration of multiple modalities through cross-attention allows the model to explore non-linear synergies between spectral, meteorological, and phenological data, surpassing late fusion or ensemble approaches.

Spatial Scale: The 10m resolution (resampled from 20m bands) represents an optimized compromise between spatial detail and computational efficiency, being adequate for the average scale of Portuguese parcels.

4.2 Implementation Considerations

Computational Efficiency: Optimization for NVIDIA RTX 4050 GPU through mixed precision training, gradient checkpointing, and adaptive batch sizes. Implementation of early stopping based on specific agricultural metrics. It is estimated that the study area covers about 40,000 hectares, totaling 4,000,000 pixels and approximately 15,625 patches.

Interpretability: Integration of attention visualization and grad-CAM for automatic generation of importance maps, allowing agronomic validation of model decisions and identification of critical periods.

Operationalization: Modular architecture allowing real-time inference with new Sentinel-2 data, including automatic preprocessing pipeline and alert generation based on validated agronomic criteria.

The proposed methodology represents an innovative contribution by combining state-of-the-art techniques with specialized agronomic knowledge, creating an operational system for modernizing Portuguese agriculture through intelligent and proactive monitoring.