Kaunas University of Technology

Faculty of Informatics

Estimation of crop growth stages from satellite images

Master's Final Degree Project

Oleksandr Laptiev

Project author

Assoc. prof. Mantas Lukoševičius

Supervisor

Kaunas, 2026

Kaunas University of Technology

Faculty of Informatics

Estimation of crop growth stages from satellite images

Master's Final Degree Project

Artificial Intelligence in Computer Science (6211BX007)

Oleksandr Laptiev

Project author

Assoc. prof. Mantas Lukoševičius

Supervisor

Prof. dr. Gintaras Palubeckis

Reviewer

Kaunas, 2026

Kaunas University of Technology

Faculty of Informatics

Oleksandr Laptiev

Estimation of crop growth stages from satellite images

Declaration of Academic Integrity

I confirm the following:

1. I have prepared the final degree project independently and honestly without any violations of the copyrights or other rights of others, following the provisions of the Law on Copyrights and Related Rights of the Republic of Lithuania, the Regulations on the Management and Transfer of Intellectual Property of Kaunas University of Technology (hereinafter – University) and the ethical requirements stipulated by the Code of Academic Ethics of the University;

2. All the data and research results provided in the final degree project are correct and obtained legally; none of the parts of this project are plagiarised from any printed or electronic sources; all the quotations and references provided in the text of the final degree project are indicated in the list of references;

3. I have not paid anyone any monetary funds for the final degree project or the parts thereof unless required by the law;

4. I understand that in the case of any discovery of the fact of dishonesty or violation of any rights of others, the academic penalties will be imposed on me under the procedure applied at the University; I will be expelled from the University and my final degree project can be submitted to the Office of the Ombudsperson for Academic Ethics and Procedures in the examination of a possible violation of academic ethics.

Oleksandr Laptiev

Confirmed electronically

Kaunas University of Technology

Faculty of Informatics

Estimation of crop growth stages from satellite images

Topic of the project	Estimation of crop growth stages from satellite images
Requirements and conditions	
Supervisor	assoc. prof. Mantas Lukoševičius

(position, name, surname, signature of the supervisor) (date)

Oleksandr Laptiev. Estimation of crop growth stages from satellite images. Master's Final Degree Project. supervisor Assoc. prof. Mantas Lukoševičius; Faculty of Informatics, Kaunas University of Technology.

Study field and area (study field group): Computer Science, Informatics (B01)

Keywords: Artificial Intelligence, Precision agriculture, Growth stage.

Kaunas, 2026. 64 pages.

Summary

This research project addresses the estimation of crop growth stages on the BBCH scale from satellite data as a regression task. Ground-truth measurements collected with the Smart Agrometer for corn (14 observations, BBCH 10–51) and winter wheat (BBCH 21–69, two growing seasons) were combined with Sentinel-1 SAR backscatter (VH, VV) and Sentinel-2 multispectral reflectance via the Copernicus Data Space API, with a ±2-day temporal tolerance and spatial averaging over field polygons. Fourteen input features were produced, including raw spectral bands and three vegetation indices (NDVI, NDWI, RVI). Three models of increasing complexity were trained and compared: Ridge regression, a feed-forward neural network (FFNN), and a Long Short-Term Memory (LSTM) network, with hyperparameters tuned using Bayesian optimization. For corn, the LSTM achieved an MSE of 0.23 on an independent test set, against 13.37 for the FFNN and 70.03 for Ridge regression. For winter wheat, evaluated by 6-fold GroupKFold cross-validation with field-level separation, the LSTM reached $R^2 = 0.639$, RMSE = 8.18, and MAE = 5.45, outperforming the FFNN ($R^2 = 0.451$) and Ridge ($R^2 = 0.205$). Permutation Importance, SHAP, and LIME revealed crop-specific feature hierarchies: NDVI and NDWI dominated for corn, while RVI together with near-infrared and red-edge bands dominated for winter wheat. An XAI-driven feature selection experiment (reducing 14 features to 9) improved the FFNN (R^2 0.451 → 0.520) but slightly degraded the LSTM. The main limitations are the small dataset size and the underrepresentation of late-season stages (BBCH ≥ 60), which led to systematic underestimation at advanced maturation phases. The results confirm that sequence-based deep learning is the most suitable approach for continuous BBCH regression from fused radar and optical satellite data.

Oleksandr Laptiev. Pasėlių augimo tarpsnio įvertinimas iš palydovinių vaizdų. Magistro studijų baigiamasis projektas / vadovas doc. Mantas Lukoševičius; Kauno technologijos universitetas, Informatikos fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Informatikos mokslai, Informatika (B01).

Reikšminiai žodžiai: Dirbtinis intelektas, Tikslioji žemdirbystė, Augimo etapas.

Kaunas, 2026. 64 p.

Santrauka

Šiame magistro baigiamajame projekte pasėlių augimo tarpsnio įvertinimas pagal BBCH skalę iš palydovinių duomenų formuluojamas kaip regresijos uždavinys. „Smart Agrometer" įrenginiu surinkti etaloniniai matavimai kukurūzams (14 stebėjimų, BBCH 10–51) ir žieminiams kviečiams (BBCH 21–69, du vegetacijos sezonai) buvo apjungti su „Sentinel-1" SAR atgalinės sklaidos duomenimis (VH, VV) bei „Sentinel-2" daugiakanaliais atspindžio duomenimis, naudojant Copernicus Data Space API su ±2 dienų laiko tolerancija ir erdviniu vidurkinimu pagal lauko poligonus. Iš viso suformuoti 14 įvesties požymių, įskaitant spektrines juostas ir tris augmenijos indeksus (NDVI, NDWI, RVI). Apmokyti ir palyginti trys vis sudėtingėjantys modeliai: Ridge regresija, tiesioginio sklidimo neuroninis tinklas (FFNN) ir LSTM tinklas; hiperparametrai parinkti Bajeso optimizavimu. Kukurūzams LSTM nepriklausomame teste pasiekė MSE = 0.23, lyginant su 13.37 (FFNN) ir 70.03 (Ridge). Žieminiams kviečiams, vertinant 6 dalių GroupKFold kryžmine validacija atskiriant laukus, LSTM pasiekė $R^2 = 0.639$, RMSE = 8.18 ir MAE = 5.45, pranokdamas FFNN ($R^2 = 0.451$) ir Ridge ($R^2 = 0.205$). Paaiškinamojo dirbtinio intelekto metodai (Permutation Importance, SHAP, LIME) atskleidė skirtingas požymių hierarchijas: kukurūzuose dominavo NDVI ir NDWI, o žieminiuose kviečiuose – RVI kartu su artimosiomis infraraudonosiomis ir raudonojo krašto juostomis. Papildomas eksperimentas su XAI grįsta požymių atranka (sumažinant nuo 14 iki 9) pagerino FFNN rezultatus (R^2 nuo 0.451 iki 0.520), tačiau šiek tiek pablogino LSTM. Pagrindiniai darbo apribojimai – nedidelė duomenų rinkinio apimtis ir nepakankamas vėlyvųjų augimo stadijų (BBCH ≥ 60) atstovavimas, dėl ko vėlyvosiose brandos fazėse stebimas sistemingas nuvertinimas. Rezultatai patvirtina, kad sekomis grįsti gilieji modeliai yra tinkamiausi tolydžiam BBCH regresavimui iš sujungtų radarinių ir optinių palydovinių duomenų.

Table of contents

List of figures

List of tables

List of abbreviations and terms

Abbreviations:

AI – Artificial Intelligence

CNN – Convolutional Neural Network

EVI – Enhanced Vegetation Index

FFNN – Feedforward Neural Network

Fmask – Function of Mask

GBDT – Gradient Boosted Decision Trees

LIME – Local Interpretable Model-agnostic Explanations

LSTM – Long Short-Term Memory Network

MAE – Mean Absolute Error

MSE – Mean Squared Error

NDVI – Normalized Difference Vegetation Index

NDWI - Normalized difference water index

RF – Random Forest

RMSE – Root Mean Squared Error

RNN – Recurrent Neural Network

RVI – Radar Vegetation Index

SAR – Synthetic Aperture Radar

SCL – Scene Classification Layer

SHAP – SHapley Additive exPlanations

SVM – Support Vector Machines

SVR – Support Vector Regression

UAV – Unmanned Aerial Vehicle

WKT – Well-Known Text

XAI – Explainable Artificial Intelligence

Terms:

BBCH-scale – A standardized system to describe the growth stages of plants, ranging from germination to senescence, used in agriculture and plant sciences.

VH Polarization – the radar sends a signal with vertical polarization but receives the backscattered signal with horizontal polarization. VH polarization is more sensitive to vegetation structure, because vegetation tends to depolarize the radar signal as it scatters through leaves, stems, and branches.

VV Polarization – the radar sends a signal with vertical polarization and also receives the backscattered signal with vertical polarization. VV polarization is sensitive to smooth surfaces such as water bodies, bare soil, and urban areas.

Introduction

Accurate estimation of crop growth stages is a critical task in precision agriculture, providing the basis for well-grounded decisions about when to fertilize, irrigate, treat against pests and harvest. Historically, tracking how crops develop has required agronomists to walk through fields and record observations by hand, a process that scales poorly across large agricultural regions. The rapid advancement of satellite-based remote sensing technologies — particularly ESA's Copernicus programme through its Sentinel-1 radar and Sentinel-2 optical missions — now permits large-scale cropland observation at low cost using multispectral and radar data. However, what remains difficult is converting this stream of Earth-observation data into reliable, per-field BBCH estimates — a problem that calls for combining heterogeneous inputs and modern learning algorithms.

Project novelty and relevance

The relevance of this work is driven by the growing demand for automated, scalable crop monitoring systems that can support precision agriculture across diverse environmental conditions. While many prior works have used satellite imagery together with ML methods to estimate harvest volumes and assign crops to general phenological categories, relatively few have addressed the direct prediction of numeric BBCH growth stage values as a regression task at the individual field level. The present work brings together radar observations from Sentinel-1, optical imagery from Sentinel-2, and in-field BBCH readings collected with the Smart Agrometer, and uses them to train and benchmark several model families targeting continuous growth-stage estimation.

The novelty of the project lies in the regression-based formulation of BBCH estimation from fused radar and optical satellite features, combined with a comprehensive explainability analysis using three independent XAI methods (Permutation Importance, SHAP, and LIME). Unlike traditional classification approaches that assign discrete phenological labels, the proposed method predicts continuous BBCH values, providing more granular information about crop development. Furthermore, the XAI-driven feature selection experiment demonstrates how model interpretability can be used to guide practical dimensionality reduction, with model-dependent outcomes that have not been systematically documented in prior work.

Research object and problem

The object of this research is the estimation of crop growth stages from satellite-derived features using machine learning models. The problem addressed is the absence of reliable methods for predicting continuous BBCH-scale values at the individual field level from multi-source satellite data. The mapping from spectral and radar features to a numeric BBCH value is nonlinear, crop-dependent, and sensitive to the distribution of training data. Existing approaches predominantly treat phenology detection as a classification task with coarse categorical labels or focus on yield estimation, leaving a gap in methods that provide granular, continuous BBCH estimates suitable for field-level decision-making.

Aim and objectives

This research aims to build and assess machine learning architectures capable of inferring crop development phases directly from Earth observation data. The following objectives are formulated to accomplish this overarching goal:

1. Review and analyze existing scientific approaches for crop growth stage estimation using remote sensing data and machine learning methods.

2. Design a data collection and preprocessing pipeline that integrates Sentinel-1 and Sentinel-2 satellite observations with ground-truth BBCH measurements from the Smart Agrometer.

3. Implement and train three AI model architectures of increasing complexity (Ridge Regression, FFNN, LSTM) for BBCH growth stage prediction of corn and winter wheat.

4. Evaluate model performance using cross-validation with field-level separation and compare the results across architectures and crops.

5. Apply explainable AI techniques to interpret model decisions and conduct a feature selection experiment based on XAI results.

Document structure

The thesis extends over three chapters. The first chapter presents a literature review analyzing existing satellite technologies, preprocessing methods, vegetation indices, machine learning algorithms, and recent studies on crop phenology estimation from remote sensing data. The second chapter describes the methodology, including the requirements for the AI models, the tools and libraries used, the data collection and preprocessing pipeline, the model architectures, training strategies, and evaluation metrics. The third chapter presents the experimental evaluation, including dataset preparation, model training results for corn and winter wheat, XAI analysis, the feature selection experiment, and a comparative results analysis against published studies. Finally, the conclusions summarize the key findings and identify the limitations of the study.

Use of Artificial Intelligence Tools

During the preparation of this project, generative AI tool (Claude by Anthropic) supported the author with literature exploration, language polishing, debugging of scripts and document formatting. Any output produced with its help was checked, corrected where necessary, and incorporated only after the author's own judgement. The experimental design, data collection, model implementation, and interpretation of results were performed independently by the author.

1. Literature review on satellite-based crop growth stage estimation

This section reviews the relevant research studies, highlighting their main contributions, methods, and findings. The focus is on satellite and sensor usage, satellite data pre-processing, methods for determining plant growth stages, and AI model training algorithms.

1.1. Satellites and sensors

The studies primarily rely on satellite imagery and sensors for agricultural monitoring and plant growth analysis. The commonly used satellites and sensors include:

Sentinel-2 – Appears in the majority of the reviewed studies on satellite-based crop monitoring, owing to its 10–60 m spatial resolution, 5-day revisit cycle, and thirteen-channel multispectral coverage. The literature most often uses its bands to derive vegetation indicators such as NDVI (Normalized Difference Vegetation Index) and EVI (Enhanced Vegetation Index) for assessing crop health and growth stages [1, 2, 3, 4, 5, 6]. The mission's MSI sensor records thirteen channels spread across the visible, red-edge, NIR and SWIR portions of the spectrum, with pixel sizes of 10 m, 20 m or 60 m depending on the band. Notably, the red-edge bands (705–783 nm) have been identified as particularly important for phenological studies, as they are highly sensitive to chlorophyll content and canopy structure changes during different growth stages [7, 8]. Several studies have demonstrated that the 5-day revisit time of the Sentinel-2 constellation enables dense temporal sampling critical for tracking rapid phenological transitions [9].

Sentinel-1 carries a SAR instrument whose active microwave signal penetrates cloud cover, making it valuable for crop observation in regions where optical imagery is frequently obscured. Sentinel-1 data is often combined with optical satellites like Sentinel-2 to enhance accuracy in plant growth monitoring [1, 4, 10].

Landsat-8 – Frequently used for medium-resolution agricultural monitoring and historical trend analysis due to its long-term data availability. Landsat-8 imagery is often pre-processed for radiometric corrections and georeferencing before extracting vegetation indices [5, 11, 12].

MODIS (Moderate Resolution Imaging Spectroradiometer) – Known for its high temporal resolution, MODIS is suitable for monitoring large-scale vegetation changes. It provides continuous observations that are particularly helpful for tracking seasonal plant growth [12, 13, 14].

WorldView-2/3 – High-resolution commercial satellites providing multispectral and panchromatic imagery. These satellites were used to monitor detailed crop structures and growth stages due to their very fine spatial resolution. WorldView-2/3 data proved particularly useful for analyzing smaller agricultural plots and detecting fine-scale variations in vegetation indices [3].

UAV-Based Sensors – High-resolution UAV (Unmanned Aerial Vehicle) sensors, including hyperspectral and multi-spectral cameras, are often used to complement satellite data for localized field monitoring. These sensors provide detailed imagery that enhances the precision of plant growth stage detection [5, 15].

Table 1 shows a comparison of satellite capabilities, detailing their imaging capacities, sensor types, and potential applications in agriculture. The table highlights the unique strengths of each satellite,

including spatial resolution, spectral coverage, and specific use cases such as vegetation monitoring, crop health assessment, and precision agriculture under various environmental conditions.

Table 1. Comparison of satellite capabilities

Satellite	Capabilities	Special Features	Spectral Bands and Resolutions
Sentinel-2	High-resolution multi-spectral imagery, effective for vegetation monitoring and detection of indices like NDVI and EVI.	High spatial resolution; effective for detecting vegetation indices.	13 spectral bands: Coastal aerosol (442.7 nm, 60 m), Blue (492.4 nm, 10 m), Green (559.8 nm, 10 m), Red (664.6 nm, 10 m), Vegetation red edge (704.1 nm, 20 m; 740.5 nm, 20 m; 782.8 nm, 20 m), NIR (832.8 nm, 10 m), Narrow NIR (864.7 nm, 20 m), Water vapour (945.1 nm, 60 m), SWIR-Cirrus (1373.5 nm, 60 m), SWIR (1613.7 nm, 20 m; 2202.4 nm, 20 m).
Sentinel-1	Radar-based imaging with Synthetic Aperture Radar (SAR), providing all-weather monitoring capabilities.	Effective in regions with frequent cloud cover; combines well with Sentinel-2 for enhanced accuracy.	C-band SAR: Central frequency 5.405 GHz (~5.6 cm wavelength). Modes and resolutions: Interferometric Wide (5 m x 20 m, 250 km swath), Extra-Wide Swath (20 m x 40 m, 400 km swath), Stripmap (5 m x 5 m, 80 km swath).
Landsat-8	Medium-resolution imagery for agricultural monitoring and historical trend analysis; includes pre-processed data for radiometric corrections.	Long-term data availability; supports radiometric corrections and georeferencing.	Bands: Coastal/Aerosol (0.435–0.451 µm, 30 m), Blue (0.452–0.512 µm, 30 m), Green (0.533–0.590 µm, 30 m), Red (0.636–0.673 µm, 30 m), NIR (0.851–0.879 µm, 30 m), SWIR1 (1.566–1.651 µm, 30 m), SWIR2 (2.107–2.294 µm, 30 m), Panchromatic (0.503–0.676 µm, 15 m), Cirrus (1.363–1.384 µm, 30 m), TIRS1 (10.60–11.19 µm, 100 m resampled to 30 m), TIRS2 (11.50–12.51 µm, 100 m resampled to 30 m).
MODIS	High temporal resolution imaging suitable for large-scale vegetation monitoring and seasonal plant growth tracking.	Continuous observations; effective for monitoring vegetation changes over large areas.	36 spectral bands: Bands 1–2 (0.62–0.87 µm, 250 m), Bands 3–7 (0.48–2.155 µm, 500 m), Bands 8–36 (0.405–14.385 µm, 1,000 m).
WorldView-2/3	Very high-resolution multispectral and panchromatic imagery for detailed analysis of small plots and fine-scale vegetation variations.	Ultra-fine spatial resolution; excellent for detecting fine-scale variations in agricultural fields.	8 multispectral bands: Coastal (400–450 nm, 1.84 m), Blue (450–510 nm, 1.84 m), Green (510–580 nm, 1.84 m), Yellow (585–625 nm, 1.84 m), Red (630–690 nm, 1.84 m), Red Edge (705–745 nm, 1.84 m), NIR1 (770–895 nm, 1.84 m), NIR2 (860–1,040 nm, 1.84 m). Panchromatic (450–800 nm, 0.46 m).WorldView-3 has additional shortwave infrared (SWIR) bands.

1.2. Pre-processing of satellite data

Pre-processing of satellite data is a critical step in ensuring accuracy and consistency for plant growth monitoring and phenological analysis. The reviewed studies employed various pre-processing techniques to enhance the quality of satellite images and prepare them for subsequent analysis. Below is a detailed overview of the main pre-processing steps identified in the studies:

Radiometric and Atmospheric Corrections – Radiometric corrections are applied to adjust the satellite data for sensor errors and inconsistencies in reflectance values. This ensures that the data accurately represents surface reflectance, eliminating errors introduced during data acquisition. Atmospheric corrections remove distortions caused by atmospheric conditions (e.g., aerosols, clouds, water vapor) to improve spectral integrity. Tools like Sen2Cor for Sentinel-2 and MODIS Atmospheric Correction Algorithm are commonly used [11].

Georeferencing and Orthorectification – Georeferencing aligns satellite imagery to known geographic coordinates, ensuring spatial accuracy. This step is particularly essential for combining data from multiple sources like Sentinel-2, Landsat-8, and UAV sensors. Orthorectification corrects geometric distortions caused by terrain variations, satellite sensor angles, and Earth curvature. This ensures that the satellite images are spatially accurate and can be overlaid with field data or other geospatial layers [11].

Noise Reduction and Cloud Masking – Cloud cover is a major challenge in optical satellite data. Techniques like cloud masking are applied to remove or filter out cloud-contaminated pixels. For Sentinel-2 data, algorithms such as the Fmask (Function of Mask) and Scene Classification Layer (SCL) are used to identify and mask clouds. Noise reduction methods, such as spatial and spectral filtering, are applied to reduce sensor noise and enhance image clarity, especially in high-resolution imagery like WorldView-2/3 and UAV-based data [3, 15].

Data Fusion – Combines data from multiple satellites and sensors to improve spatial, temporal, and spectral resolution. Sentinel-1 (SAR) and Sentinel-2 (optical) are often fused to overcome cloud interference and obtain both spectral and radar information. UAV-based high-resolution data is fused with satellite imagery (e.g., Sentinel-2 or WorldView-3) to enhance localized precision and monitor smaller agricultural fields in detail [3, 4].

Resampling and Reprojection – Resampling standardizes the spatial resolution of imagery when combining datasets from different sensors. For example, Sentinel-2 data (10m) may be resampled to match the resolution of Landsat-8 (30m) or fused with UAV data at higher resolutions [11]. Reprojection ensures that all imagery uses the same coordinate reference system, enabling spatial consistency across datasets.

Temporal Compositing and Gap Filling – Due to the limited revisit frequency of individual satellite missions and frequent cloud contamination, continuous time series of vegetation observations are rarely available from a single sensor. Temporal compositing techniques address this by selecting the best available pixel within a defined time window (e.g., maximum NDVI compositing over 8–16 day periods), thereby reducing cloud artifacts while preserving phenological signals. More advanced approaches employ harmonic fitting or Savitzky-Golay filtering to reconstruct smooth, gap-free vegetation index trajectories from irregularly spaced observations. These reconstructed time series serve as the foundation for both threshold-based and model-based phenology extraction methods [9].

Speckle Filtering for SAR Data – Because radar pulses are coherent, SAR images carry an intrinsic grainy noise known as speckle. It degrades the usable signal in backscatter measurements and propagates errors into index calculations. Common speckle reduction techniques include the Lee filter, Refined Lee filter, and Gamma-MAP filter, all of which balance noise suppression with preservation of spatial detail. For Sentinel-1 GRD products, multi-temporal filtering approaches are particularly effective, as they exploit the temporal redundancy of repeat-pass acquisitions to average out random speckle while retaining the systematic backscatter changes associated with crop growth [16, 17]. In the context of this work, Sentinel-1 data was preprocessed using standard ground range detected products with terrain correction applied.

Spectral Band Selection and Feature Engineering – Not all spectral bands contribute equally to crop monitoring tasks. Band selection is often guided by prior knowledge of crop spectral responses at different growth stages. For instance, Sentinel-2 red-edge bands (B05, B06, B07) are particularly sensitive to chlorophyll content and canopy structure changes during vegetative growth, while shortwave infrared bands (B11, B12) respond to leaf water content and are informative during senescence and maturation stages. The process of deriving vegetation indices (such as NDVI, NDWI, EVI, and RVI) from raw spectral bands represents a form of feature engineering that compresses multi-band information into biophysically meaningful indicators. The choice of which indices to compute and which raw bands to retain as model inputs is a design decision that directly affects model performance [8, 9].

1.3. Methods for determining the stage of plant growth and important indicators

The reviewed studies employed a variety of methods to determine plant growth stages and identify critical indicators. These methods primarily rely on spectral indices, temporal analysis, and integration of field data with satellite-derived information.

Spectral indices derived from multi-spectral and hyperspectral satellite data are widely used to determine plant growth stages. These indices measure the reflectance of vegetation in specific spectral bands (e.g., red, near-infrared) and provide insights into crop health and phenology. Key indices include:

Normalized Difference Vegetation Index is the de-facto standard index for monitoring crop vitality and developmental phases, computed from the normalized gap between near-infrared and red reflectance values. Elevated NDVI readings indicate dense, photosynthetically active canopies [1, 3, 11]. Studies have confirmed that NDVI time-series from Sentinel-2 can effectively capture phenological transitions such as tillering, heading, and maturity in cereal crops [9]. However, NDVI tends to saturate in dense vegetation canopies, limiting its sensitivity during peak growth periods [18].

Enhanced Vegetation Index was designed to suppress soil-brightness and atmospheric interference that affect NDVI, yielding cleaner estimates [2]. The Plant Phenology Index has also been proposed as an alternative to NDVI and EVI, particularly demonstrating advantages in evergreen forests and snow-covered areas [18].

Soil Adjusted Vegetation Index accounts for soil reflectance in sparse vegetation regions, where bare soil can affect the accuracy of NDVI [3, 13].

Green Chlorophyll Index is used to monitor chlorophyll content in crops, which is a critical indicator of plant health and nitrogen status [13].

Radar Vegetation Index (RVI) derived from Sentinel-1 SAR data has gained significant attention for crop phenology monitoring. RVI utilizes the ratio of cross-polarization (VH) to total backscatter (VV+VH), and its temporal evolution correlates well with the main phenological stages of crops [16, 17, 19]. Research has demonstrated that SAR-derived vegetation descriptors can effectively assess crop growth processes for wheat, corn, canola, and soybean, with the cross-polarization ratio showing particular sensitivity to changes in canopy biomass and plant water content [17]. Recent work by Mandal et al. proposed the DpRVI descriptor which outperforms traditional backscatter ratios and simple RVI for correlations with crop biophysical parameters [19].

Time-series analysis involves monitoring spectral indices over a growing season to detect changes in plant health and identify phenological events. Methods include:

Curve Fitting and Breakpoint Detection: vegetation index curves (e.g., NDVI curves) are fitted to identify breakpoints corresponding to key growth stages such as onset of greening, peak growth, and harvest [12]. A recent phenology extraction framework based on the Derivative Dynamic Time Warping algorithm demonstrated the ability to simultaneously extract complete BBCH-scale phenological stages for corn from Sentinel-2 NDVI time series, achieving RMSE values of less than 6 days across all detected stages [20].

Spectral Index Calculation – After pre-processing, indices like NDVI, EVI and SAVI are derived to characterize canopy condition and follow developmental progress. Spectral index extraction involves calculating ratios between bands (e.g., red, near-infrared) to highlight specific plant growth indicators. The calculated indices are often used as inputs for machine learning models and time-series analysis [3, 13].

Temporal Aggregation: satellite images are aggregated into weekly or monthly intervals to reduce noise and improve trend detection. Sentinel-1, Sentinel-2 and MODIS data are often used for this purpose due to their high temporal resolution [4, 5].

Phenology Metrics: derived from time-series data, metrics like "start of season", "peak of season" , and "end of season" are used to characterize plant growth stages and detect anomalies caused by environmental stressors.

Near Real-Time (NRT) Phenology Detection: Liao et al. proposed an NRT phenology framework to detect and forecast BBCH-scale phenology for winter wheat and corn from Sentinel-2 time-series data. This approach directly estimates the BBCH stage on a specific date rather than detecting only transition dates, making it particularly relevant for crop management decisions. The framework uses shape model fitting combined with accumulated growing degree days to achieve sub-field level phenology monitoring [21].

Machine Learning-Based Phenology Detection – Beyond traditional threshold and curve-fitting approaches, machine learning methods have been increasingly applied to phenological stage detection. Ensemble tree learners — Random Forest in particular — have been deployed to assign time-stacked satellite observations to specific phenological labels, leveraging the ensemble learning capability to handle noisy and high-dimensional input features. Gradient-boosting variants such as

XGBoost and LightGBM tend to perform competitively, especially when paired with a preliminary variable-importance step that identify the most discriminative spectral and temporal variables. These approaches typically treat phenology detection as a classification problem, assigning each observation to one of several predefined growth stages [7, 22].

Deep Learning for Temporal Pattern Recognition – Deep learning architectures designed for sequential data processing have opened new possibilities for phenology estimation. Time-aware deep models (RNNs and their gated descendants like LSTM and GRU) are well-suited to learning when phenological transitions occur from vegetation-index series. Spatial deep models (CNNs) operate on raster tiles and surface within-field variation through learned local features. More recently, hybrid architectures combining CNNs with LSTMs or attention mechanisms have been proposed to jointly capture spatial and temporal features, achieving state-of-the-art performance in crop type classification and phenological stage prediction [23, 24].

Regression-Based BBCH Estimation – A fundamentally different approach to phenology estimation is to treat the problem as a regression task, directly predicting the numeric BBCH value rather than classifying discrete stages or detecting transition dates. This formulation provides continuous, granular information about crop development and can leverage the ordinal nature of the BBCH scale. However, regression-based approaches require training data with precise BBCH annotations and face challenges related to the nonlinear mapping between spectral features and growth stage values. The work of Liao et al. is among the few that adopt this regression formulation for BBCH estimation from Sentinel-2 time series. [21].

1.4. Studies on crop growth stage prediction using remote sensing and machine learning

A growing body of research has focused specifically on predicting crop growth stages from satellite imagery using various machine learning and deep learning approaches. This section reviews key studies whose results can be compared with those obtained in this work.

Phenology-based classification of winter wheat using Sentinel-2 imagery was investigated by Newete et al., who applied the Random Forest algorithm to classify winter wheat during different phenological stages. The study found that the crop 'heading' stage yielded the highest producer's accuracy of 92.52%, while the overall classification accuracy reached 83.58% during the maturity stage. NDVI values extracted from Sentinel-2 time series and grouped using K-means clustering were effective in distinguishing different phenological phases [22].

Jiao et al. investigated the integration of Sentinel-1 SAR and Sentinel-2 optical imagery with a crop structure dynamics model for tracking crop condition. Their research demonstrated that the SAR-derived vegetation condition was strongly correlated with crop biomass early in the season (R^2 = 0.88). The combination of radar and optical features proved particularly effective for capturing phenological transitions during leaf development, stem elongation, and flowering stages [25].

Nasirzadehdizaji et al. investigated the correlation between Sentinel-1 interferometric coherence, backscatter values, and different phenological stages of maize, sunflower, and wheat. Their results demonstrated a strong relationship between SAR products and crop growing seasons, with coherence values highest after field preparation and decreasing sharply as vegetation developed. This confirmed that multi-temporal SAR analysis can serve as an effective tool for crop growth monitoring [17].

Tian and colleagues proposed an LSTM-based predictor that fuses satellite observations with weather variables to refine wheat yield estimates. Their LSTM model captured both the variation trend and the temporal dependence structure of time-series satellite data, demonstrating superior performance compared to traditional regression approaches. The study highlighted that the internal time-step parameter of the LSTM significantly influenced prediction accuracy, and that incorporating remote sensing features from specific growth stages improved overall model performance [23].

Yang et al. estimated wheat yield using multiple machine learning approaches (LSTM, RF, GBDT, SVR) based on Sentinel-2 multispectral and ZY-1 02D hyperspectral data. The LSTM model achieved the best performance with an RMSE of 0.201 t/ha, outperforming Random forest (RMSE = 0.260 t/ha), Gradient Boosted Decision Trees (RMSE = 0.306 t/ha), and Support Vector Regression (RMSE = 0.489 t/ha). The study also revealed that 10-meter Sentinel-2 data outperformed 30-meter data, highlighting the importance of spatial resolution in crop monitoring [26].

Joshi et al. tackled winter-wheat yield forecasting with an interpretable Bi-LSTM design using remote sensing and weather data across the USA from 2008–2021. The study employed SHAP and other XAI techniques to interpret model decisions, finding that vegetation indices during the grain-filling period (corresponding to late BBCH stages) had the greatest impact on yield prediction. The Bi-LSTM achieved better accuracy than both standard LSTM and Random Forest models, demonstrating the value of bidirectional sequence processing for capturing crop growth dynamics [24].

Dhillon et al. demonstrated that integrating Random Forest with crop growth models improved winter wheat and oil seed rape yield predictions. Their research showed that NDVI alone was insufficient for accurate predictions ($R^2 < 0.65$ for wheat), but combining NDVI with climate parameters improved model performance ($R^2 > 0.70$). The study emphasized that variable selection and the combination of multiple data sources play critical roles in achieving accurate crop predictions [27].

Meroni and co-authors derived per-crop phenological signals from joint Sentinel-1/Sentinel-2 time series, covering nine widespread European crops and using LUCAS field observations as reference. They compared start-of-season (SOS), peak-of-season (PS), and end-of-season (EOS) timings derived from NDVI and the SAR cross-ratio (CR) against observed BBCH stages from the German Weather Service (DWD). For wheat, the Sentinel-1-derived SOS and EOS timings showed deviations of 4.7 days for early stem elongation (BBCH 31) and 5.7 days for late ripening (BBCH 87). They also found that VV polarization was generally more correlated with NDVI than VH, and that the temporal evolution of coherence fitted well the main phenological stages for most crop types [28].

An important consideration across all model training approaches is the integration of multi-source satellite data as input features. Mateo-Sanchis, Anna et al. [14] examined how optical and microwave channels complement each other when estimating yields and demonstrated that combining Sentinel-1 SAR and optical remote sensing data provides complementary information that neither source can offer alone. Optical data captures chlorophyll absorption, canopy greenness, and water stress indicators through vegetation indices, while microwave data penetrates cloud cover and responds to canopy structure, biomass, and soil moisture. Their findings showed that models trained on fused optical-microwave inputs consistently outperformed those relying on a single data modality, confirming that the information content of SAR and multispectral observations is largely complementary rather than redundant. The choice of loss function and optimization strategy also significantly influences model training outcomes. For regression tasks such as BBCH prediction, the

mean squared error (MSE) loss function is the most commonly used objective, as it penalizes larger prediction errors more heavily and provides smooth gradients for optimization. The Adam and AdamW optimizers have become the default choice for training neural networks in agricultural applications due to their adaptive learning rate mechanism, which automatically adjusts the step size for each parameter based on the history of gradients. AdamW separates the weight-decay term from the gradient step, which yields cleaner regularization than the implicit L2 penalty baked into Adam. Hyperparameter tuning is a critical but often underreported aspect of model training. The performance of neural network models is highly sensitive to design parameters (depth, width, dropout) and the optimizer schedule (step size, mini-batch length, epoch count). Grid search and random search are simple but computationally expensive strategies, as the number of evaluations grows exponentially with the dimensionality of the search space. Bayesian search replaces uninformed sweeps with a Gaussian-process surrogate that estimates the loss landscape and proposes candidate configurations with the highest expected improvement. This approach has been successfully applied in crop monitoring studies to tune LSTM and FFNN architectures, reducing the computational burden while identifying better-performing configurations than manual tuning.

These studies collectively demonstrate the growing capability of remote sensing and machine learning techniques for crop phenology monitoring and growth stage prediction. However, most existing research focuses on either yield prediction or broad phenological transition detection, rather than direct BBCH-scale estimation at the field level. The present work contributes to this emerging area by developing models that directly predict numeric BBCH values from integrated Sentinel-1 and Sentinel-2 features.

1.5. Model training algorithms

Model training algorithms play a key role in processing satellite and field data to analyze plant growth stages, classify vegetation, and predict phenological events. The reviewed studies employed various approaches for training models, ranging from traditional statistical methods to advanced machine learning techniques. Below is a detailed overview of the model training algorithms and their implementation in relevant studies.

Convolutional Neural Networks are widely used for spatial feature extraction and classification tasks in high-resolution satellite and UAV imagery. The deep learning capabilities of CNNs allow for automatic identification of crop structures and phenological stages. Multi-spectral imagery from satellites such as Sentinel-2 and WorldView-3, as well as hyperspectral UAV data, were used as input data. Detecting growth stages, identifying vegetation stress, and segmenting agricultural fields. For example, in one study, high-resolution data from WorldView-3 and UAV sensors were processed using CNNs to detect subtle growth variations in crops [2, 15].

Random Forest is one of the most frequently used algorithms in remote-sensing phenology because it tolerates noisy, high-dimensional inputs without extensive feature engineering. In the reviewed literature it is applied both to discrete-stage classification and to regression-style estimation, with input vectors typically built from Sentinel-2, Landsat-8 or MODIS-derived spectral indices (NDVI, EVI, SAVI). The ensemble averaging that defines the method also tempers overfitting on small or imbalanced training sets — a recurring concern in field-level phenology work. Time-series NDVI fed into a Random Forest has, for instance, been used to separate germination, flowering and maturity transitions across the growing season [3, 10].

Support Vector Machines are commonly used for crop classification and phenological stage detection, particularly when datasets are small or require binary classification. Spectral indices such as NDVI and SAVI, as well as field-validated observations, were used as input. SVMs effectively identify crop types, detect growth stages, and monitor vegetation health. One study demonstrated the application of SVMs for classifying phenological stages of wheat using Sentinel-2 data [11].

LSTM networks belong to the recurrent family and were originally introduced to overcome the vanishing-gradient problem that limits plain RNNs on long sequences. In the context of crop monitoring, the natural input is a temporally ordered stack of spectral indices — usually NDVI from Sentinel-2, MODIS or Landsat-8 — and the natural target is a phenological milestone such as start-of-season, peak-of-season or end-of-season. The gating mechanism is what makes the architecture attractive here: it lets the network learn which parts of the index trajectory are diagnostic of a transition and which are seasonal noise. De Freitas Cunha and Silva, for example, applied an LSTM to multi-season NDVI series and reported clear gains over linear and tree-based regression baselines in capturing growth-stage dynamics [29].

Linear Regression sits at the base of the statistical toolkit and is used whenever one needs a transparent mapping between explanatory inputs and a continuous response. In the context of agricultural monitoring, Linear Regression has been applied to relate spectral indices (such as NDVI or EVI) to biophysical parameters, including biomass, leaf area index, and yield estimates. Its simplicity and interpretability make it a valuable baseline for comparison against more complex models. For example, several studies used Linear Regression to estimate crop growth metrics from Sentinel-2 derived indices, providing quick assessments of vegetation status and temporal trends across growing seasons [3, 10].

Feedforward Neural Networks form a class of neural architectures characterized by acyclic information flow: each layer receives activations only from preceding layers. They consist of multiple layers of interconnected neurons, allowing them to model nonlinear relationships between input features and target variables. In agricultural applications, FFNNs have been employed to integrate multi-source data such as satellite imagery, weather parameters, and soil characteristics to predict phenological stages, estimate yields, or classify crop types.

Bidirectional LSTM (Bi-LSTM) augments the unidirectional LSTM by running two parallel passes over the input — left-to-right and right-to-left — so each step receives information from observations on either side. This allows the model to capture both past and future context for each time step, which can be beneficial for phenological analysis where the trajectory of vegetation indices before and after a given date both carry information about the current growth stage. Joshi, Abhasha et al. demonstrated that Bi-LSTM outperformed standard LSTM for winter wheat yield prediction ($R^2 = 0.81$ vs. $R^2 = 0.78$), and used SHAP to identify vegetation indices during mid-to-late growth stages as the most influential features [24].

Explainable AI (XAI) in agricultural modeling has gained increasing attention as researchers seek to understand not only what models predict but why they make specific predictions. In the context of crop growth stage estimation, XAI techniques serve multiple purposes: validating that models rely on physically meaningful features, identifying crop-specific spectral signatures, and building trust in model predictions for practical agricultural applications. The three most commonly used XAI methods in agricultural remote sensing are SHAP (SHapley Additive exPlanations), which provides

theoretically grounded feature attribution based on cooperative game theory; LIME (Local Interpretable Model-agnostic Explanations), which creates local linear approximations of model behavior; and Permutation Importance, which measures the degradation in prediction accuracy when feature values are randomly shuffled. The combined application of multiple XAI techniques, as employed in this work, provides a more comprehensive and robust assessment of feature relevance than any single method alone [24].

1.6. Analytical methods and instruments

This section introduces the analytical instrument and the satellite data sources selected for this study on the basis of the literature reviewed in the previous sections. Subsection 1.6.1 describes the Smart Agrometer field device used to collect the ground-truth BBCH measurements, and Subsection 1.6.2 presents the Sentinel-1 and Sentinel-2 spectral and radar bands chosen as model inputs.

1.6.1. Smart agrometer

The Smart Agrometer is a handheld device designed to assist agronomists in diagnosing nutrient deficiencies directly in the field. When using the device, an agronomist inputs the type of crop and the plant's growth stage, the instrument records the coordinates and time of measurement. It measures leaf spectra within the 400–850 nm range, requiring a minimum leaf width of 5 mm, and completes each measurement in approximately 4 seconds. For accurate diagnostics, it is recommended to sample 20 plants per session. The device features GSM and GPS connectivity, ensuring precise location tracking, and offers a battery life of up to 24 hours on a full charge. After data collection, results are processed through a cloud-based algorithm, providing insights within minutes via a web-based analysis tool. Developed and tested over 15 years, the Smart Agrometer achieves laboratory-level standards, with over 85% of its deficiency diagnoses aligning with nutrient concentration tests. Its compact and user-friendly design, integrated with mobile software, requires minimal training, enabling agronomists to make informed fertilization decisions by accurately identifying nutrient limitations in crops at specific times and locations. Figure 1 shows what this device looks like. [30]

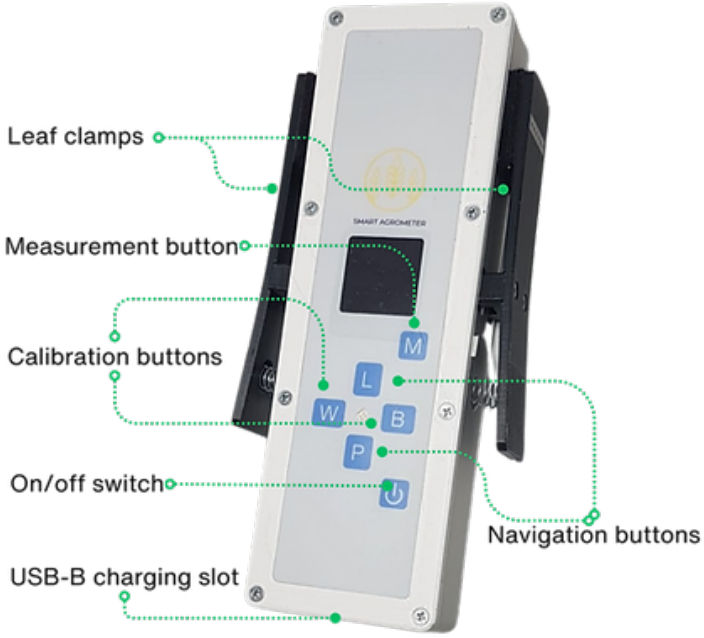

Fig. 1. Smart agrometer

1.6.2. Selected data sources for model training

Data from the Sentinel-1 and Sentinel-2 satellites and the Smart Agrometer were used as inputs to train machine learning models for predicting plant growth stages from satellite imagery. Sentinel-1 provides C-band radar data through two polarizations, VV and VH, both with a resolution of 10 meters. These radar bands are particularly useful for monitoring soil moisture and plant structure, even under cloud cover. Sentinel-2 offers a range of spectral bands across different wavelengths, providing detailed information on vegetation health and growth. Band 1 (Coastal aerosol) at 0.443 µm and Band 2 (Blue) at 0.490 µm support atmospheric corrections and water-body detection, while Band 3 (Green) at 0.560 µm and Band 4 (Red) at 0.665 µm are essential for assessing chlorophyll content in plants. Bands 5, 6, and 7, all in the Red Edge spectrum at 0.705 µm, 0.740 µm, and 0.783 µm respectively, are crucial for monitoring plant stress and biomass changes. The Near Infrared (NIR) Band 8 at 0.842 µm and the Narrow NIR Band 8A at 0.865 µm are key indicators of vegetation health and density. Band 9, measuring water-vapour absorption at 0.945 µm, assists in atmospheric corrections, while Band 10 (SWIR-Cirrus) at 1.375 µm is used to detect high-altitude clouds. Bands 11 and 12 in the Short-Wave Infrared (SWIR) range, at 1.610 µm and 2.190 µm respectively, are valuable for assessing leaf moisture content and distinguishing between different crop types. The combination of radar and optical data from these bands was used alongside field measurements from the Smart Agrometer, which provided ground-truth data on crop type, growth stage, and field coordinates with timestamps. This integration of satellite imagery and ground-based measurements

made it possible to train models predicting plant growth stages, which in turn reduces the number of in-field measurements that agronomists need to take. Table 2 summarises the radar channels and optical bands used in each acquisition, together with their native ground sampling distance (GSD) and the role each plays in the feature engineering pipeline of this study.

Table 2. Spectral and radar bands used in this study

Mission	Band ID	Spectral region	Central λ / f	GSD (m)	Role in feature vector
Sentinel-1	VV	Microwave (C-band, co-pol)	5.404 GHz	10	RVI denominator
	VH	Microwave (C-band, cross-pol)	5.404 GHz	10	RVI numerator, canopy structure
Sentinel-2	B01	Coastal aerosol	0.443 µm	60	Excluded (atmospheric correction only)
	B02	Visible — Blue	0.490 µm	10	Used as raw band
	B03	Visible — Green	0.560 µm	10	NDWI numerator
	B04	Visible — Red	0.665 µm	10	NDVI denominator term
	B05	Red-edge (lower)	0.705 µm	20	Chlorophyll-sensitive feature
	B06	Red-edge (mid)	0.740 µm	20	Chlorophyll-sensitive feature
	B07	Red-edge (upper)	0.783 µm	20	Chlorophyll-sensitive feature
	B08	NIR (broad)	0.842 µm	10	NDVI/NDWI numerator term
	B08A	NIR (narrow)	0.865 µm	20	Canopy density proxy
	B09	Water vapour	0.945 µm	60	Used as raw band
	B10	SWIR — Cirrus	1.375 µm	60	Excluded (cloud detection only)
	B11	SWIR (lower)	1.610 µm	20	Leaf-water content
	B12	SWIR (upper)	2.190 µm	20	Leaf-water content, senescence

2. Methodology for BBCH growth stage prediction from satellite imagery

This section provides a comprehensive description of the methodology employed in this study, including the requirements for the AI models, the tools and libraries used, the data collection and preprocessing pipeline, the model architectures and training strategies, the hyperparameter optimization procedure, the evaluation framework, and the explainability methods applied to interpret model decisions.

2.1. Formation of requirements

The AI model for predicting crop growth stage should meet the following requirements.

Training data sources: The model should use data from multiple sources for training, including satellite imagery from Sentinel-1 (VV, VH radar bands) and Sentinel-2 (optical bands from visible to shortwave infrared), as well as ground measurements from Smart Agrometer, which provide information on crop type, growth stage, and geolocation coordinates with timestamps.

Input requirements: The model receives satellite observations at multiple wavelengths, field coordinates, crop type, and observation date as input. Specifically, the input feature vector consists of Sentinel-2 spectral band values (B02–B12, excluding B01 and B10 due to their low spatial resolution and atmospheric correction purpose) and three derived vegetation indices (NDVI, NDWI, and RVI). The RVI index is computed from Sentinel-1 radar data (VV and VH backscatter coefficients), which are not included as standalone input features. All spectral values and derived indices are spatially averaged over each field polygon to produce a single representative value per field and observation date.

Model output: The model should predict the growth stage of crops on the BBCH-scale, formulated as a regression task. The BBCH scale is a standardized system describing plant development stages numerically, ranging from 10 to 90. The model outputs a continuous numeric BBCH value, which can then be rounded to the nearest principal growth stage for practical interpretation.

Pre-processing requirements: The model should include data harmonization and pre-processing steps such as cloud masking for Sentinel-2 images, atmospheric corrections, temporal alignment of satellite observations with ground-truth measurements using a ± 2-day buffer window to account for satellite pass frequency and cloudiness, and standard normalization of all input features to improve model convergence and training stability.

Accuracy and performance: The model should achieve high accuracy in predicting crop growth stages, as measured by mean squared error (MSE) and root mean squared error (RMSE) between predicted and observed BBCH values. The model should generalize well across different spatial locations, which is ensured through field-level train/validation splitting strategies.

Spatial generalization: The model must be able to predict BBCH growth stages for agricultural fields that were not included in the training set. This requirement is critical for practical deployment, where the model is applied to new fields without prior observations. To enforce this constraint, field-level separation was maintained throughout all data splitting procedures, ensuring that no field appeared simultaneously in training and validation/test partitions.

Interpretability: The model predictions should be interpretable through established explainable AI techniques. This requirement serves two purposes: (1) validating that the model relies on physically meaningful spectral features rather than spurious correlations, and (2) providing agronomists with insight into which satellite-derived indicators are most informative for different crops and growth stages, potentially guiding future data collection efforts.

2.2. Description of tools

This subsection describes the software tools and libraries used throughout the experimental pipeline, from data collection through model training, evaluation, and interpretation.

Python was chosen as the primary programming language due to its extensive ecosystem for scientific computing, machine learning, and geospatial data processing. All experiments were conducted in Jupyter Notebook environments to facilitate interactive development and reproducibility.

PyTorch was used as the deep learning framework for developing and training the Feed-Forward Neural Network (FFNN) and Long Short-Term Memory (LSTM) models. PyTorch offers a dynamic computational graph, which allows models to be defined and modified during runtime, providing maximum flexibility in architecture design. GPU acceleration via CUDA was leveraged to speed up training. Specific PyTorch modules used include: nn.LSTM for recurrent layers; nn.Linear for fully connected layers; nn.LayerNorm for layer normalization; nn.Dropout for regularization; and nn.MSELoss as the loss function. The AdamW optimizer (optim.AdamW) was used for all neural network training, combining adaptive learning rates with decoupled weight decay regularization.

scikit-learn was employed for several purposes: (1) implementing the Ridge regression model (a linear regression variant with L2 regularization) as a baseline; (2) computing evaluation metrics including mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2); (3) implementing the GroupKFold cross-validation strategy, which ensures that samples from the same agricultural field are never split between training and validation sets; and (4) performing StandardScaler normalization, which standardizes features by removing the mean and scaling to unit variance, fitted on the training set and applied consistently to validation and test sets to prevent data leakage.

scikit-optimize (skopt) was used for Bayesian hyperparameter optimization of the FFNN and LSTM models. The BayesSearchCV function performs hyperparameter search using Gaussian process-based Bayesian optimization, which is more sample-efficient than grid or random search. The search spaces included hidden layer sizes, number of LSTM layers, and dropout rates, with negative mean squared error as the scoring metric.

rasterio was used for reading and processing geospatial raster data from satellite imagery stored in GeoTIFF format. This library enabled the extraction of pixel values from individual spectral bands for each field polygon, supporting the computation of spatial averages and vegetation indices. It was also used to read and display true-color composite images for visual validation of model predictions.

Copernicus Data Space API was used for the collection of satellite data, providing programmatic access to Earth observation products from the Sentinel-1 and Sentinel-2 missions. This API enables efficient querying, downloading, and management of large-scale geospatial datasets. Its flexible

search parameters and compatibility with standard data formats (such as WKT Polygon) make it possible to collect all the necessary satellite information required for this work. Queries were configured with specific parameters including the field polygon in WKT format, observation date with a ± 2-day tolerance window, cloud cover thresholds (for Sentinel-2), and specific band selections for each satellite mission.

SHAP (SHapley Additive exPlanations) was used as an explainability tool based on cooperative game theory. It assigns each feature a contribution value (Shapley value) reflecting its marginal impact on the prediction. The shap.Explainer was applied to all three model types, using the training data as the background dataset and computing SHAP values for each validation sample. For the cross-validated winter wheat models, SHAP values were computed per fold and then aggregated (mean and standard deviation) across all folds.

LIME (Local Interpretable Model-agnostic Explanations) was used to provide localized, linear approximations of model behavior around individual predictions. The LimeTabularExplainer was configured in regression mode and applied to each model. LIME generates perturbations of input features and fits a local linear model to approximate the decision boundary, enabling insight into how specific features influenced individual outcomes. As with SHAP, for the cross-validated wheat experiments, LIME explanations were computed for each validation sample across all folds and then averaged.

Permutation Importance was implemented as a custom function that measures changes in prediction performance (MSE) when individual feature values are randomly permuted. For each feature, the original values were shuffled multiple times (n_repeats = 3), and the resulting increase in MSE compared to the baseline was recorded. Features whose permutation caused the largest degradation in model performance were deemed most important. This approach is model-agnostic and was applied identically to all three model architectures.

pandas and NumPy were used throughout the project for data manipulation, numerical computations, and managing the ground-truth dataset (including parsing observation dates, field identifiers, GPS coordinates, and BBCH growth stage labels).

matplotlib and seaborn were used for all data visualization, including growth stage distributions, vegetation index trends, loss curves, feature importance bar charts, SHAP summary plots, LIME explanation plots, and spatial prediction maps overlaid on true-color satellite imagery.

2.3. Data collection and preprocessing pipeline

This subsection provides a comprehensive description of the data acquisition, preprocessing, and feature engineering steps applied to construct the training dataset from raw satellite imagery and ground-truth observations.

2.3.1. Ground-truth data collection

Ground-truth measurements were obtained using the Smart Agrometer device and manual agronomist observations conducted during the 2024 and 2025 growing seasons. Each observation record contains the field identifier, GPS coordinates (latitude and longitude), a timestamp indicating the date of measurement, the crop type (corn or winter wheat), and the recorded growth stage on the BBCH scale. Field boundaries were defined as polygons using Well-Known Text (WKT) representations, which

ensured accurate spatial correspondence between ground measurements and satellite imagery. The agronomists visited fields at approximately weekly intervals during active growth periods, with more frequent visits during rapid phenological transitions.

For corn, observations were collected from 6 distinct agricultural fields, yielding 14 total measurements spanning BBCH stages 10 (leaf development) through 51 (beginning of heading). For winter wheat, data was collected from a larger set of fields across two growing seasons, providing approximately 60 observations covering BBCH stages 21 (beginning of tillering) through 69 (end of flowering). The temporal distribution of measurements was uneven, with denser sampling during early-to-mid season stages and sparser coverage at late-season stages, particularly for BBCH values above 60.

2.3.2. Satellite data acquisition

Satellite imagery was collected using the Copernicus Data Space API, which provides programmatic access to Earth observation products from the Sentinel-1 and Sentinel-2 missions. For each ground-truth observation, the system queried satellite data within a ±2-day temporal window centered on the observation date to account for the satellite revisit frequency and cloud cover conditions. This buffer was necessary because the ground measurement date does not always coincide with a cloud-free satellite acquisition.

Sentinel-2 Level-2A products (atmospherically corrected surface reflectance) were downloaded for bands B02 through B12, excluding B01 (Coastal aerosol, 60 m resolution) and B10 (SWIR-Cirrus, 60 m resolution) due to their low spatial resolution and primarily atmospheric correction purposes. The excluded bands do not carry meaningful vegetation information at the field scale. The remaining 11 bands cover the blue (B02, 490 nm), green (B03, 560 nm), red (B04, 665 nm), red-edge (B05 at 705 nm, B06 at 740 nm, B07 at 783 nm), near-infrared (B08 at 842 nm, B8A at 865 nm), water vapor (B09 at 945 nm), and shortwave infrared (B11 at 1610 nm, B12 at 2190 nm) spectral regions. Sentinel-1 Ground Range Detected (GRD) products were downloaded for VV (co-polarization) and VH (cross-polarization) channels, providing radar backscatter information independent of weather and illumination conditions.

2.3.3. Feature extraction and spatial aggregation

For each satellite observation matched to a ground-truth record, the following processing steps were applied. First, each spectral band (Sentinel-2) or polarization channel (Sentinel-1) was read as a geospatial raster using the rasterio library. Second, pixel values falling within the field polygon boundary were extracted. Third, invalid values (NaN, caused by cloud masking or sensor artifacts) were replaced with zeros using NumPy's *nan_to_num* function. Fourth, the spatial mean of all valid pixels within the polygon was computed to produce a single representative value per band per field and date. This spatial averaging approach reduces within-field variability and produces a compact feature representation suitable for the machine learning models used in this work.

A temporal matching function (*find_closest_date*) was implemented to pair each ground-truth observation with the closest available satellite acquisition within a ±3-day tolerance window. If no satellite data was available within this window (e.g., due to persistent cloud cover), the observation was excluded from the dataset. This function iterates through the directory structure of downloaded satellite data, parses acquisition dates from folder names, and selects the temporally nearest match.

2.3.4. Vegetation index computation

Three vegetation indices were computed from the extracted band values to enhance the dataset with biophysically meaningful variables:

The Normalized Difference Vegetation Index exploits the contrast between strong NIR reflectance from healthy leaf mesophyll and red-light absorption by chlorophyll:

$$NDVI = \frac{B08 - B04}{B08 + B04}$$

Output values fall in the closed interval [−1, +1]; for the BBCH range covered by the corn and winter wheat datasets, observed NDVI mostly falls between 0.2 (bare or sparsely covered soil at early vegetative stages) and 0.7 (closed canopy at heading and flowering). Because of this dynamic range and its sensitivity to chlorophyll content, NDVI is included as one of three derived features in the input vector.

A second derived feature, NDWI, captures canopy water status through the green–NIR contrast:

$$NDWI = \frac{B03 - B08}{B03 + B08}$$

Within the present dataset, NDWI typically takes negative values (between −0.5 and −0.3) and shifts downward as the canopy approaches grain filling — a behaviour that complements NDVI, which tends to saturate once the canopy is closed. NDWI was selected specifically to provide signal during late BBCH stages where chlorophyll-driven indices lose discriminative power.

The third derived feature is computed from the dual-polarisation Sentinel-1 acquisition:

$$RVI = \frac{4 \cdot VH}{VV + VH}$$

The rationale is structural rather than spectral. Co-polarised backscatter (VV) is dominated by surface scattering from soil and smooth canopy elements, whereas the cross-polarised channel (VH) responds to volume scattering produced by randomly oriented stems, leaves and branches. As the canopy fills in, the VH-to-VV ratio rises, pushing RVI upward independently of cloud cover — which makes the index a useful fallback during periods when Sentinel-2 acquisitions are obscured [19].

2.3.5. Feature normalization

All extracted spectral values, radar values, and computed vegetation indices were normalized using StandardScaler from the scikit-learn library. The scaler was fitted exclusively on the training partition to prevent data leakage: the mean and standard deviation of each feature were computed from training samples only, and the learned transformation was then applied identically to validation and test samples. This normalization ensures that all features operate on a comparable scale (zero mean, unit variance), which is critical for the convergence behavior of gradient-based optimization algorithms used in neural network training. Without normalization, features with larger absolute values (e.g., shortwave infrared reflectance) would disproportionately influence the loss function gradients.

2.3.6. Final feature vector and dataset statistics

The resulting feature vector for each sample consisted of 14 values: 11 Sentinel-2 spectral band means (B02, B03, B04, B05, B06, B07, B08, B8A, B09, B11, B12) and 3 derived vegetation indices (NDVI, NDWI, RVI). The target variable is the BBCH growth stage as recorded by the agronomist during the corresponding field visit.

For corn, the final dataset contained 14 valid samples distributed across 6 fields, with 9 samples assigned to training, 3 to validation, and 2 to testing. For winter wheat, the dataset contained approximately 60 valid samples distributed across multiple fields and two growing seasons (2024 and 2025), used in a 6-fold GroupKFold cross-validation scheme where the grouping variable was the field identifier.

2.4. Description of the architecture of AI models and training methods

This section describes the architectures of the artificial intelligence models used in the experiments and the corresponding training methodologies. The selected models represent different levels of complexity, ranging from classical statistical approaches to deep learning architectures capable of modeling nonlinear and temporal relationships in satellite-derived data. All models were designed to predict crop growth stages expressed on the BBCH scale using multisource remote sensing data and ground-truth measurements.

Three main model types were implemented and evaluated: Linear Regression, Feed-Forward Neural Networks (FFNN), and Long Short-Term Memory (LSTM) neural networks. Each model uses the same preprocessed input features but differs in its ability to capture nonlinearities and temporal dependencies in the data.

2.4.1. Linear Regression (Ridge)

The Linear Regression model serves as a baseline and provides a transparent and interpretable framework for analyzing relationships between spectral features and crop growth stages. Ridge regression with L2 regularization was applied to reduce overfitting and improve generalization. The L2 penalty constrains the magnitude of the model coefficients, preventing any single feature from dominating the prediction. The model was trained by minimizing the mean squared error (MSE) between predicted and observed BBCH values using the closed-form solution provided by scikit-learn's Ridge class. Due to its linear nature, this model captures only direct linear dependencies between input features and the target variable, serving as a lower-bound reference for the more complex architectures.

2.4.2. Feed-Forward Neural Network (FFNN)

An FFNN was added to the model suite specifically because linear approaches cannot represent curvature in the feature-to-phenology mapping. The network ingests the 14-dimensional descriptor, passes it through two fully connected hidden layers equipped with nonlinearities, and emits a single scalar BBCH value.

The hidden layers used Rectified Linear Unit (ReLU) activations, defined as $f(x) = max(0, x)$, which enable efficient gradient propagation while introducing the nonlinearity necessary for learning complex feature interactions. Dropout regularization was applied after each hidden layer to mitigate

overfitting, especially given the limited size of the ground-truth dataset. During training, dropout randomly sets a proportion of neuron activations to zero, effectively training an ensemble of sub-networks and reducing co-adaptation between neurons.

Optimisation was performed with AdamW (lr = 1×10^{-3}, default β parameters, decoupled weight decay) against the mean squared error between predicted and observed BBCH values. The training budget was capped at 1000 epochs, but in practice early stopping on validation MSE terminated training well before this limit. The checkpoint corresponding to the minimum observed validation error was retained as the final model state, ensuring that the reported test-set performance reflects the best-generalising rather than the last-trained configuration.

2.4.3. Long Short-Term Memory neural network

The third architecture explored in this work is built around an LSTM core. Although each ground-truth record is a single field-level observation rather than a multi-step trajectory, the 14-dimensional descriptor was treated as a length-one sequence and passed through the recurrent core, allowing the gating mechanisms to model interactions between features in a non-additive way. This design choice was motivated by the heterogeneity of the input — optical reflectances, radar backscatter ratios and derived vegetation indices respond differently across BBCH stages, and a gated unit can suppress or amplify each signal conditionally rather than weighting it linearly.

A standard LSTM cell maintains an internal cell state regulated by three multiplicative gates. The forget gate controls how much of the previous state is preserved; the input gate decides how strongly the new feature vector is written into memory; and the output gate determines how the resulting state is projected into the hidden representation passed downstream. In the context of this work, the practical consequence is that the network can learn, for example, to down-weight SWIR reflectance during early vegetative stages (when soil dominates the pixel) and increase its contribution at senescence — without an explicit phenology-aware feature.

Activations at the recurrent output were stabilised with layer normalisation, which mitigates internal covariate shift and reduces sensitivity to the learning-rate setting. Optimisation followed the same protocol as the FFNN: AdamW with lr = 1×10^{-3}, MSE objective, up to 1000 epochs, early stopping on validation MSE, with the lowest-validation-loss checkpoint retained for evaluation.

2.5. Training strategies and hyperparameter optimization

Different training strategies were applied depending on the crop type and data availability, reflecting the practical constraints imposed by the ground-truth collection process.

Early stopping strategy: For both FFNN and LSTM models, an early stopping mechanism was implemented to prevent overfitting. During training, the validation loss was monitored after each epoch, and the model state corresponding to the lowest validation MSE was saved. Training continued for up to 1000 epochs, but the final model used for evaluation was always the checkpoint with the best validation performance. This approach effectively determines the optimal number of training epochs in a data-driven manner, avoiding the need to manually specify this hyperparameter.

Batch size and data loading: Training data was organized into mini-batches of size 8 using PyTorch's DataLoader utility, with random shuffling enabled for training batches to improve stochastic gradient descent convergence. Validation and test data were loaded without shuffling to

ensure reproducible evaluation. The relatively small batch size was chosen to accommodate the limited dataset size while still providing sufficient gradient noise for effective optimization.

Reproducibility considerations: Random seeds were set for NumPy, PyTorch, and Python's built-in random module at the beginning of each experiment to ensure reproducibility of results across runs. The CUDA deterministic mode was enabled for GPU computations. All model architectures, hyperparameters, training configurations, and evaluation results were logged for each fold of the cross-validation procedure.

2.5.1. Corn training strategy

For corn, the dataset size was limited (14 total observations across 6 unique fields), so a fixed train–validation–test split was used with strict field-level separation to prevent spatial data leakage. Specifically, 4 fields were assigned to training (9 measurements), 1 field to validation (3 measurements), and 1 field to testing (2 measurements). This ensures that the model is evaluated on fields it has never seen during training, providing a realistic assessment of spatial generalization. Due to the small dataset size, hyperparameters were set manually based on preliminary experiments rather than automated optimization.

2.5.2. Winter wheat training strategy

For winter wheat, where more observations were available across multiple fields and two growing seasons (2024 and 2025), GroupKFold cross-validation was employed with K=6 folds, using field identifiers as grouping variables. This ensured that samples from the same field never appeared simultaneously in training and validation sets, providing a more realistic evaluation of model generalization across spatial locations. Each fold produced an independent model, and final predictions were obtained by averaging the outputs of all fold-specific models.

2.5.3. Bayesian hyperparameter optimization

For the winter wheat experiments, hyperparameter optimization was performed for the FFNN and LSTM models using BayesSearchCV from the scikit-optimize library. The Bayesian optimization procedure uses a Gaussian process surrogate model to estimate the objective function (negative MSE) and selects the next set of hyperparameters to evaluate by balancing exploration (trying new regions of the hyperparameter space) and exploitation (refining around promising configurations).

The search spaces were defined as follows: for the FFNN, the first hidden layer size ranged from 128 to 4096, the second hidden layer from 64 to 2048, and dropout rates from 0.1 to 0.6; for the LSTM, the number of layers ranged from 1 to 4, the hidden state size from 128 to 1024, and dropout rates from 0.1 to 0.5. Each configuration was evaluated using the GroupKFold cross-validation scheme with field identifiers as groups, ensuring spatial integrity throughout the optimization process. The best-performing hyperparameter sets were then used to train the final fold-specific models.

2.6. Evaluation metrics and model interpretability

Model performance was evaluated using several regression-based metrics:

Mean Squared Error (MSE) was the primary evaluation metric, computed as the average of the squared differences between predicted and actual BBCH values. Lower MSE indicates better model

performance. For cross-validated experiments, MSE was computed per fold and then averaged to obtain the mean and standard deviation across folds.

Root Mean Squared Error (RMSE) is the square root of MSE and provides an error estimate in the same units as the target variable (BBCH stages), facilitating direct interpretation of prediction accuracy.

Mean Absolute Error (MAE) measures the average magnitude of prediction errors without squaring, providing a robust metric less sensitive to outliers than MSE.

Coefficient of Determination (R^2) measures the proportion of variance in the observed BBCH values that is explained by the model predictions. An R^2 of 1.0 indicates perfect prediction, while values near 0 indicate that the model is no better than predicting the mean.

In addition, explainable artificial intelligence (XAI) techniques – including permutation importance, SHAP, and LIME – were applied to all trained models to analyze feature contributions and improve interpretability. For the cross-validated winter wheat models, XAI results were computed per fold and aggregated (reporting mean and standard deviation) to provide robust estimates of feature importance. These methods enabled a detailed comparison of how different architectures utilize optical, radar, and vegetation index features when predicting crop growth stages.

2.7. XAI-driven feature selection and model retraining

Following the initial training and evaluation of all models using the complete 14-dimensional feature vector, the results of the explainability analysis were used to guide a feature selection procedure aimed at reducing input dimensionality and potentially improving model generalization. The rationale for this approach is twofold. First, in the presence of limited training data, high-dimensional input spaces increase the risk of overfitting, particularly for flexible architectures such as FFNN that can memorize noise in uninformative features. Second, identifying and removing features with negligible predictive contribution simplifies the model and reduces computational cost without sacrificing — and potentially improving — prediction accuracy.

The feature selection was performed by aggregating the results of all three XAI methods (Permutation Importance, SHAP, and LIME) across all folds of the GroupKFold cross-validation for winter wheat. For each method, the mean absolute importance score was computed per feature, and the features were ranked. Features that consistently appeared in the bottom ranks across all three methods were considered candidates for removal. Cross-method consensus protects the resulting feature ranking from being driven by quirks of a particular XAI algorithm: Permutation Importance measures predictive contribution through performance degradation, SHAP provides theoretically grounded marginal attributions, and LIME captures local feature relevance around individual predictions.

Based on this analysis, the following 9 features were retained for the reduced model: six Sentinel-2 spectral bands (B06, B07, B8A, B09, B11, B12) and three derived vegetation indices (NDVI, NDWI, RVI). The removed features included visible-spectrum bands (B02, B03, B04), one red-edge band (B05) and the broadband NIR (B08). The visible-spectrum bands (blue, green, red) were found to carry minimal additional information beyond what is already captured by NDVI and NDWI, which are computed from the red, green, and NIR bands.

All three models (Ridge Regression, FFNN, and LSTM) were then retrained from scratch using only the selected 9 features, following the identical training protocols, GroupKFold cross-validation scheme (K=6), Bayesian hyperparameter optimization procedure, and evaluation metrics as in the full-feature experiments. This ensures that any differences in performance can be attributed solely to the change in input features rather than differences in training methodology. The comparison between full-feature and selected-feature models provides insight into the trade-off between input dimensionality and prediction accuracy for each architecture, and offers practical guidance for deploying growth stage prediction systems with minimal sensor requirements.

3. Experimental evaluation of crop growth stage prediction models

This section describes the dataset on which the models were trained and the results of the AI models (Linear Regression, FFNN, LSTM) for corn and winter wheat.

3.1. Dataset overview and preparation

The dataset used in this study combines two main sources: ground measurements and satellite imagery. Ground measurements are stored in a dataset that includes records of observation sessions, timestamps, GPS coordinates (latitude and longitude), field identifiers, crop types, and recorded growth stages using the BBCH scale, a standardized scale for describing plant development. Each row corresponds to a specific field observation for different crops, with dates covering the entire growing season.

To complement the field data, satellite imagery was collected using the Copernicus API, which provides access to high-resolution multispectral imagery from Sentinel-2 and radar imagery from Sentinel-1. Field polygons were defined using Well-Known Text (WKT) representations stored in the ground-truth dataset, which ensured an accurate spatial correspondence between field measurements and satellite data. Extensive preprocessing steps were performed before training the models: field coordinates were cleaned and reformatted, observation dates were aligned, and satellite imagery was extracted for time windows around each ground measurement, including a ±2 day buffer to account for satellite pass frequency and cloudiness, since the ground measurement date does not always coincide with a satellite pass and good weather conditions. Queries were written using the Copernicus API to download true color images, vegetation indices, and images in different color spectra.

After collecting the data, an analysis was conducted. Figure 2 illustrates the distribution of measurement dates alongside the corresponding recorded growth stages for corn. As observed, the available data spans growth stages BBCH 10 to BBCH 51. Notably, data for growth stages beyond BBCH 52 are currently unavailable, indicating that observations for the later phases of the growth cycle have not yet been captured by agronomists.

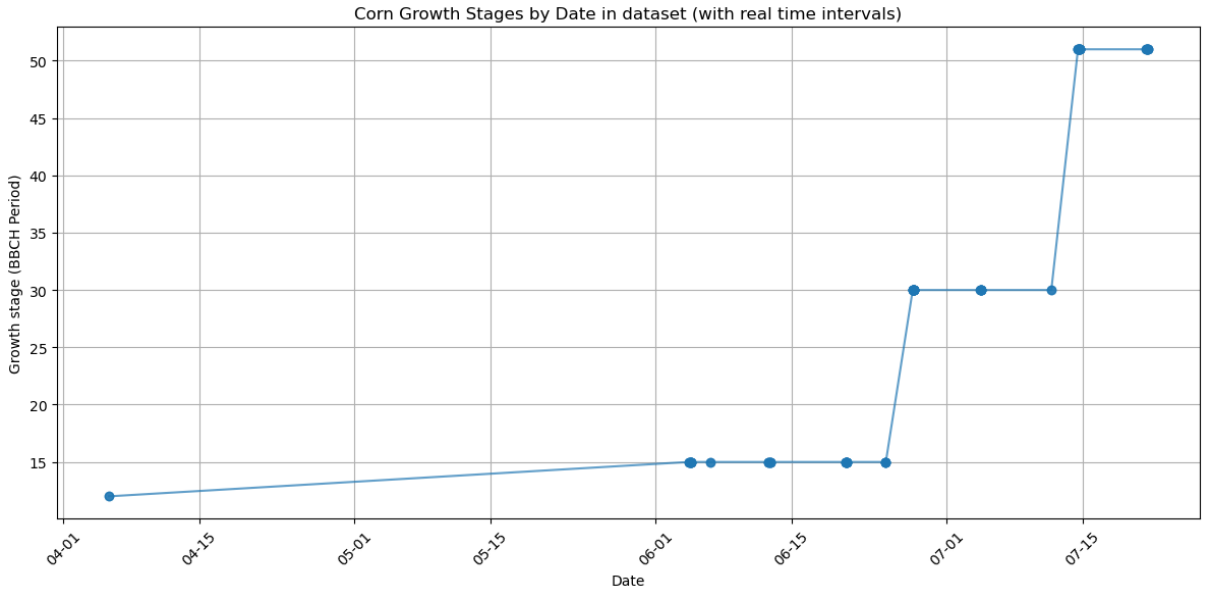

Fig. 2. Corn growth stages by date in the dataset

In addition to corn, a dataset for winter wheat was incorporated into the study. The winter wheat data was collected using the same ground-truth protocol and satellite data extraction pipeline as for corn, ensuring methodological consistency. Ground observations include field identifiers, geographical coordinates, timestamps, and BBCH growth stages specific to winter wheat phenology. The winter wheat dataset covers two growing seasons (2024 and 2025) and includes measurements primarily from BBCH 21 to 69, corresponding to stem elongation, heading, and flowering phases. Compared to corn, winter wheat exhibits a denser temporal sampling during early to mid-season stages, enabling improved temporal phenology analysis. Figure 3 illustrates the distribution of winter wheat growth stages by date across both years.

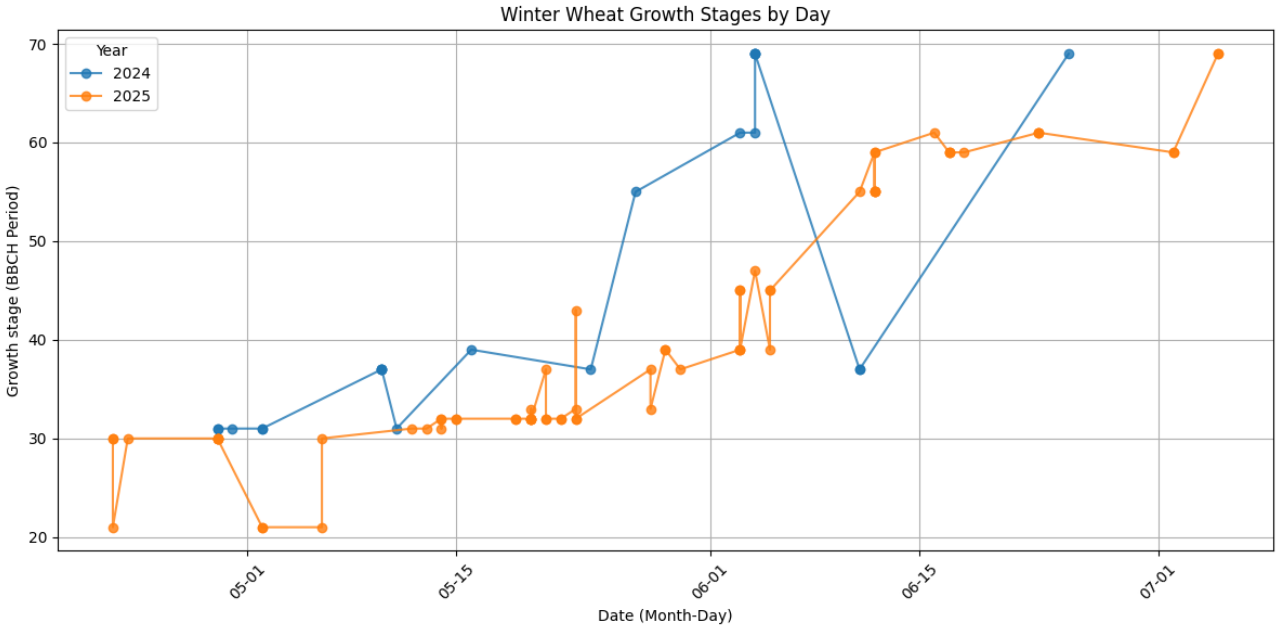

Fig. 3. Winter wheat growth stages by date in the dataset

All bands except B01 and B10 were used for model training, as these bands have low spatial resolution and are primarily intended for atmospheric correction purposes. In addition to the raw spectral bands, two widely used vegetation indices — the NDVI and the NDWI — were calculated based on Sentinel-2 bands B3 (green), B4 (red), and B8 (near-infrared), using the following formulas:

$$NDVI = \frac{B08 - B04}{B08 + B04}$$

$$NDWI = \frac{B03 - B08}{B03 + B08}$$

Furthermore, the Radar Vegetation Index (RVI) was derived from Sentinel-1 data, specifically utilizing the VV and VH polarizations, according to the formula:

$$RVI = \frac{4 \cdot VH}{VV + VH}$$

These indices were computed to enhance the dataset with biologically meaningful variables, supporting improved model performance in vegetation monitoring and phenological stage prediction tasks.

Figure 4 shows the average NDVI, NDWI, and RVI values for corn across BBCH stages, while Figure 5 presents the corresponding values for winter wheat. As shown in Figure 4, for corn, RVI and NDVI are positively correlated and both increase with advancing growth stage, while NDWI exhibits the opposite trend and decreases monotonically.

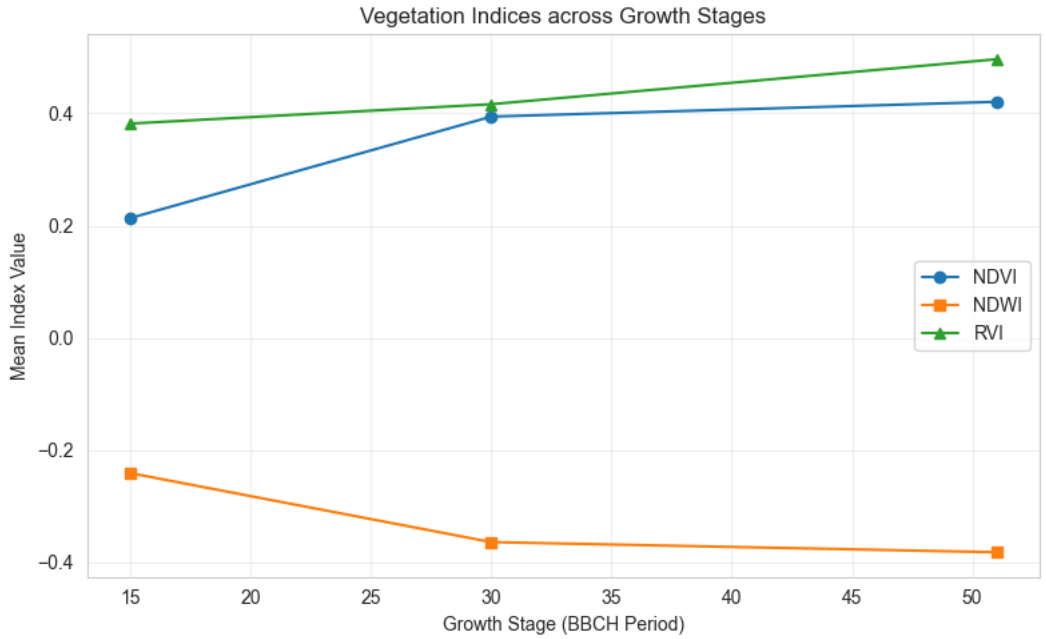

Fig. 4. Corn vegetation indices across growth stages

Figure 5 confirms that the same relationship holds for winter wheat: NDVI and RVI rise with phenological development, whereas NDWI declines and remains inversely correlated with the other two indices.

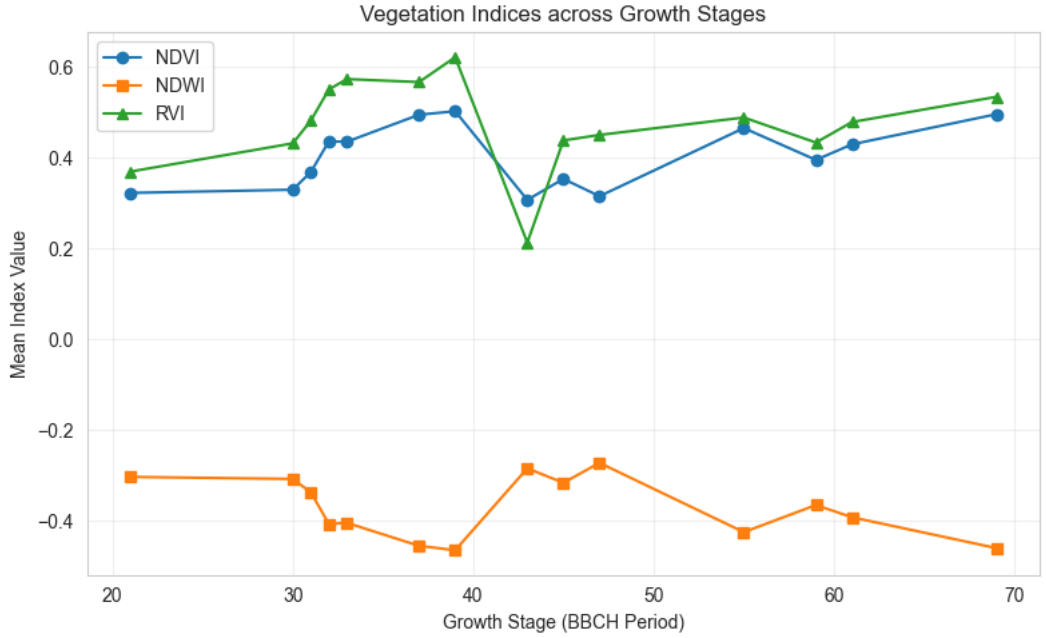

Fig. 5. Winter wheat vegetation indices across growth stages

All values extracted from the used bands were spatially averaged over each field polygon to produce a single representative value per field and date. These aggregated values were then normalized to a standard range to ensure comparability across bands and to improve the stability and convergence of the machine learning models during training.

The complete dataset was divided into training, validation, and test sets following a field-level split strategy to prevent spatial data leakage. For the corn dataset, the fields included in the training set were entirely distinct from those in the validation and test sets, ensuring no spatial overlap between them. This means that while all subsets contained only corn fields, the validation and test sets included fields that were never seen by the model during training, providing a robust evaluation framework for assessing generalization to unseen spatial locations. In total, the corn training set included 9 measurements, the validation set contained 3 measurements, and the test set comprised 2 measurements, reflecting the limited number of available ground-based observations.

For the winter wheat dataset, a different training strategy was applied due to the larger number of available observations. The model was trained using a GroupKFold cross-validation scheme with K = 6, where the grouping variable corresponded to the field identifiers. This resulted in six different train/validation splits, each ensuring that observations from the same field were never shared between training and validation sets. During testing, predictions from all six trained models were averaged to obtain the final output, providing a more stable and robust estimate of model performance for winter wheat.

3.2. Description of the created models and evaluation

This section provides an overview of the selected models, their training settings, and evaluation results, including explainable AI analyses.

3.2.1. Linear Regression for corn

The Ridge regression model, a variant of linear regression with L2 regularization, was trained using prepared dataset. Explainable AI methods, specifically SHAP, LIME and permutation importance, were employed to assess feature importance. As evidenced by Figures 6–8, the model is influenced most by the NDVI and NDWI indices, as well as bands from the infrared spectrum, which are intended to determine vegetation.

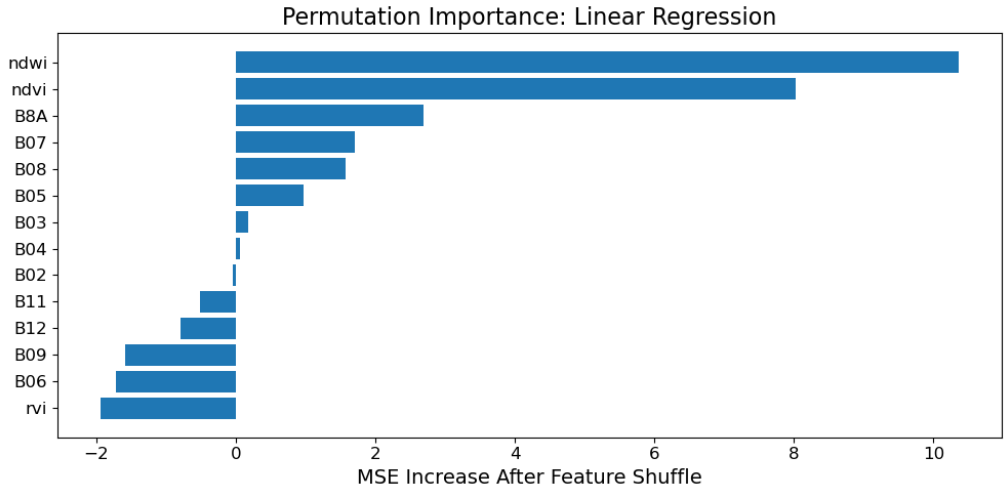

Fig. 6. Permutation Importance for Linear Regression corn model

The Permutation Importance analysis (Figure 6) reveals that NDWI produces the largest MSE increase when shuffled (~10.5), making it the most influential predictor, followed closely by NDVI (~8.0) and the narrow NIR band B8A (~2.7). Red-edge bands B07 and B08 show moderate importance (~1.5), while visible bands (B02, B03, B04) contribute minimally. Notably, several features including RVI, B06, B09, B12, and B11 show negative or near-zero importance, indicating that shuffling them actually improves or does not affect model performance — a sign that the linear model does not meaningfully rely on these inputs.

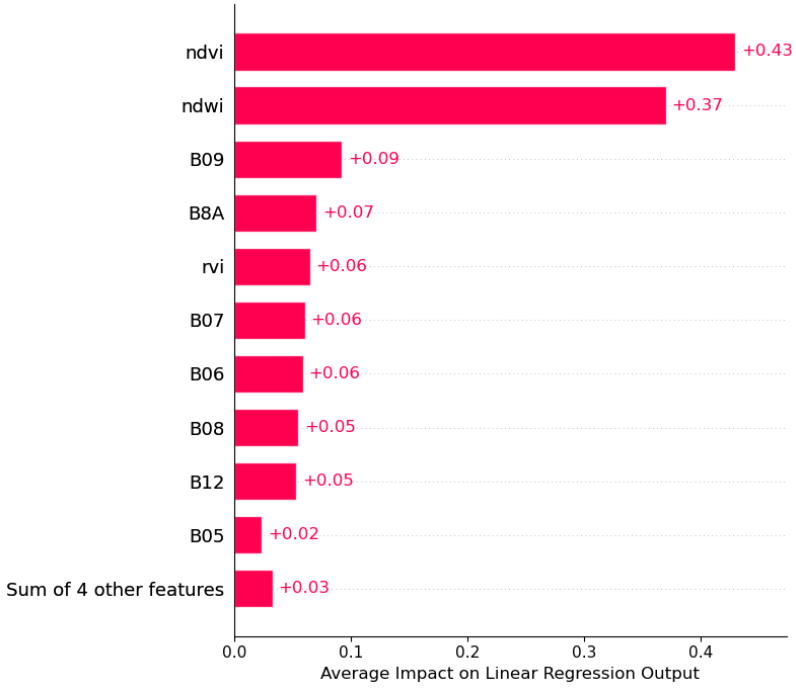

Fig. 7. SHAP explanation for Linear Regression corn model

The SHAP analysis (Figure 7) presents a different ranking: NDVI leads with the highest average impact (+0.43), followed by NDWI (+0.37), with B09 (+0.09) and B8A (+0.07) at a considerably lower level. RVI contributes only +0.06 in SHAP, consistent with its negligible permutation importance.

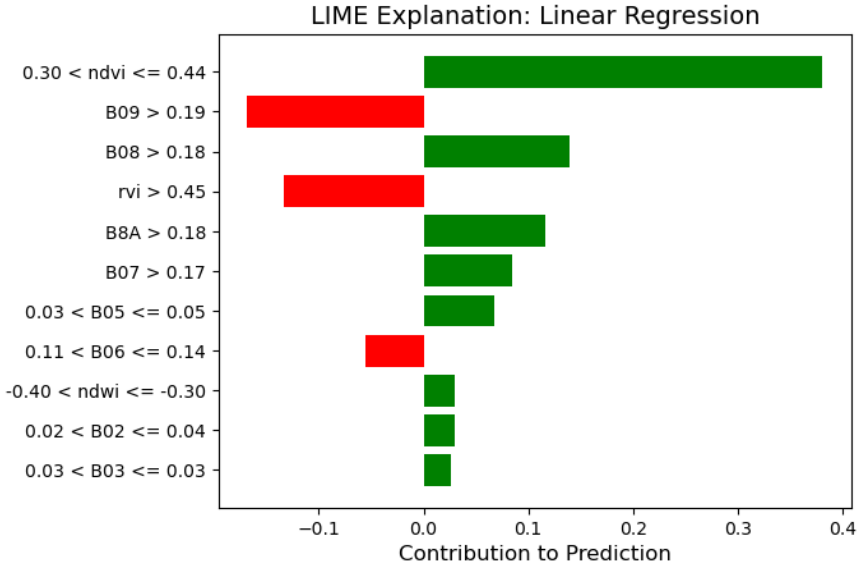

Fig. 8. LIME explanation for Linear Regression corn model

The LIME explanation (Figure 8) provides a local perspective on individual predictions, showing that NDVI in the range 0.30–0.44 has the strongest positive contribution (~0.38), while B08 and B8A also push predictions positively. Interestingly, B09 and RVI show negative LIME contributions (red bars), meaning that for specific samples their values pull the predicted BBCH downward, contrasting with their modest positive global SHAP impact.

3.2.2. Feed-forward neural network for corn

The FFNN was implemented as a multilayer perceptron comprising an input layer, two hidden layers with 256 and 128 neurons respectively, ReLU activation functions, and dropout rates of 0.2 to mitigate overfitting. The output layer consisted of a single neuron for regression. The network was trained using the AdamW optimizer with a learning rate of 0.001 and the mean squared error loss function over 1,000 epochs.

Performance evaluation using MSE showed improved accuracy compared to Ridge regression, especially when capturing nonlinear feature interactions. Analysis using XAI methods revealed similar top features to the linear model, but as can be seen from Figures 9 - 11, RVI also began to have a stronger influence on the model results.

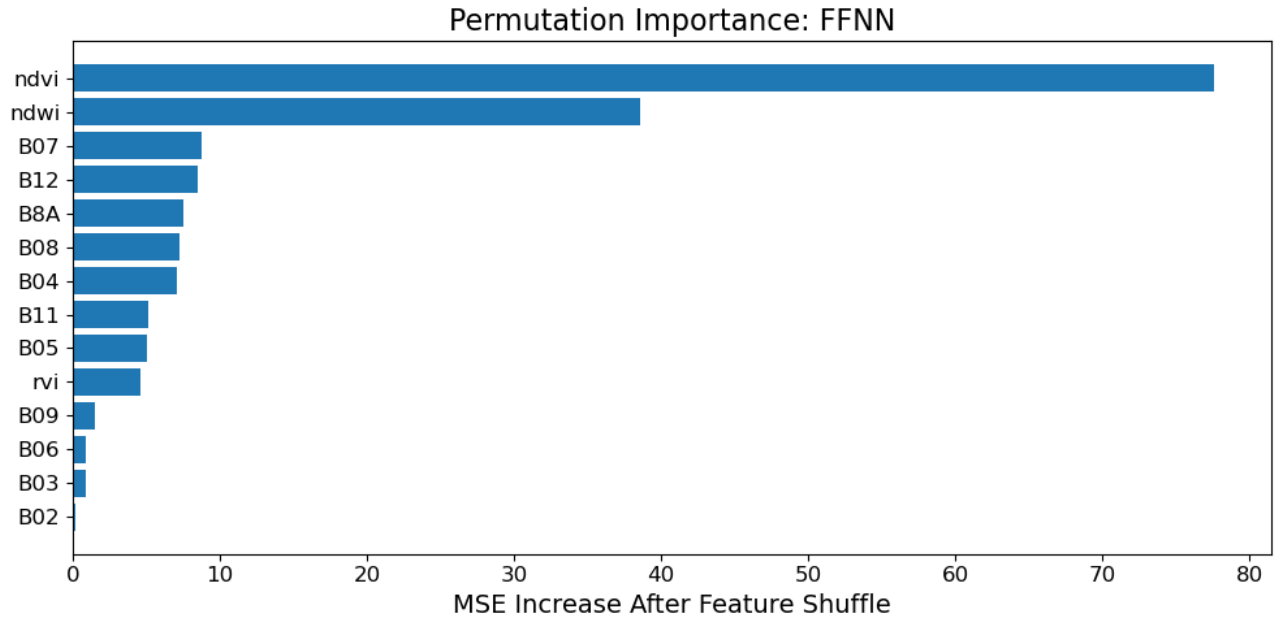

Fig. 9. Permutation Importance for FFNN corn model

The Permutation Importance analysis (Figure 9) shows that NDVI is overwhelmingly the most important feature for the FFNN, with an MSE increase of approximately 78 upon shuffling — nearly double that of NDWI (~38), which ranks second. Compared to the linear model, the FFNN distributes importance more broadly across spectral bands: B07 and B12 (~9 each), B8A and B08 (~7 each), and even the visible band B04 (~7) now show meaningful contributions. RVI remains low (~4), and B09, B06, B03, and B02 contribute minimally.

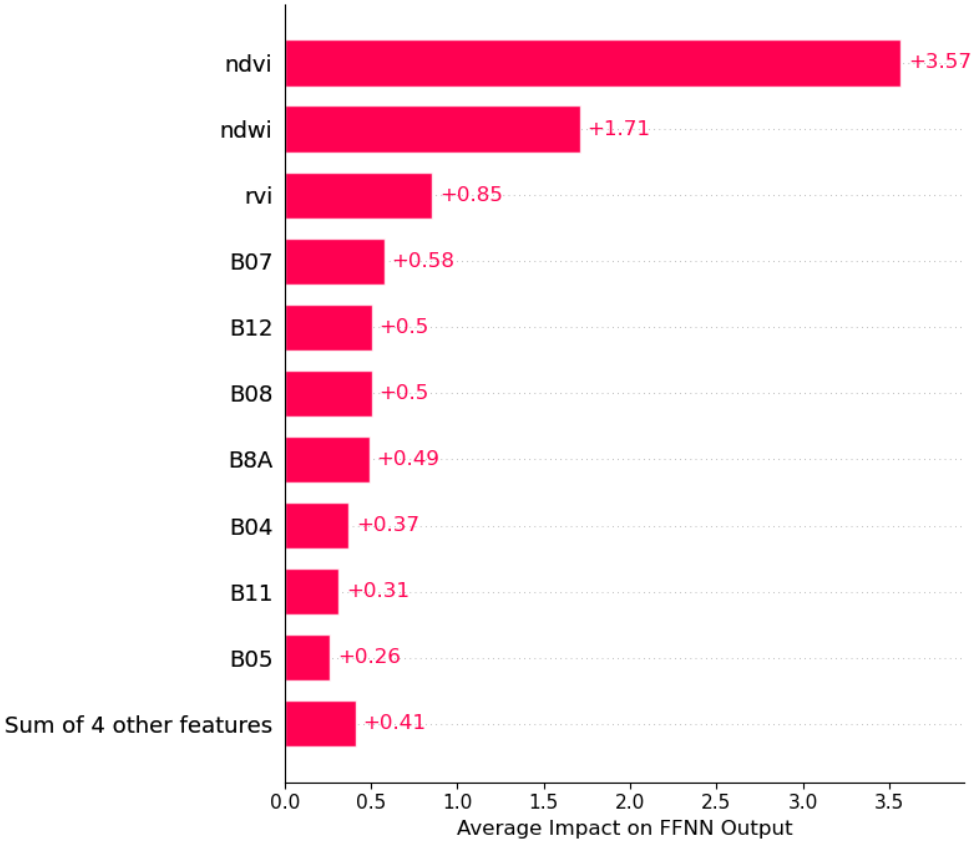

Fig. 10. SHAP explanation for FFNN corn model

The SHAP analysis (Figure 10) confirms NDVI as the dominant feature (+3.57 average impact), followed by NDWI (+1.71). Notably, RVI now ranks third in SHAP importance (+0.85), a substantial increase compared to the linear model where it scored only +0.06. B07, B12, B08, B8A, B04, B11, and B05 all show moderate SHAP values in the range +0.26 to +0.58, indicating that the FFNN exploits a wider set of spectral features than the linear baseline.

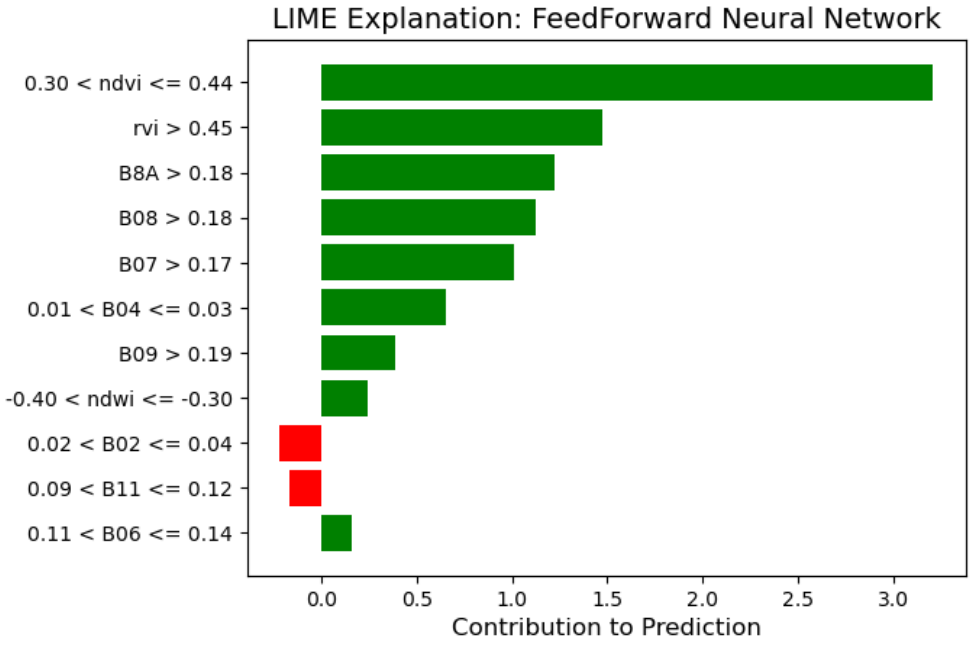

Fig. 11. LIME explanation for FFNN corn model

The LIME explanation (Figure 11) reveals that at the local prediction level, NDVI in the range 0.30–0.44 has the strongest positive contribution (~3.2), followed by RVI above 0.45 (~1.5), B8A above 0.18 (~1.3), and B08 above 0.18 (~1.2). B07 and B04 also push predictions positively. Only B02 and B11 show small negative contributions (red bars). This confirms that the FFNN captures nonlinear feature interactions that the linear model missed, particularly for RVI and red-edge bands.

3.2.3. Long short-term memory neural network for corn

For corn, the LSTM core was configured with two stacked recurrent layers of 256 hidden units and inter-layer dropout of 0.4. The recurrent output passed through layer normalisation, a 128-unit fully connected layer with ReLU, and a single linear unit emitting the predicted BBCH value. Training used the same AdamW/MSE configuration as the FFNN, ensuring that any performance difference between the two architectures can be attributed to model capacity rather than optimisation setup. The LSTM was designed to process sequential input data, capturing complex dependencies across observation periods.

Explainable AI methods, demonstrated that the LSTM predominantly leveraged complex patterns in NDVI and NDWI indices. Figures 12–14 show the complete results of the XAI methods.

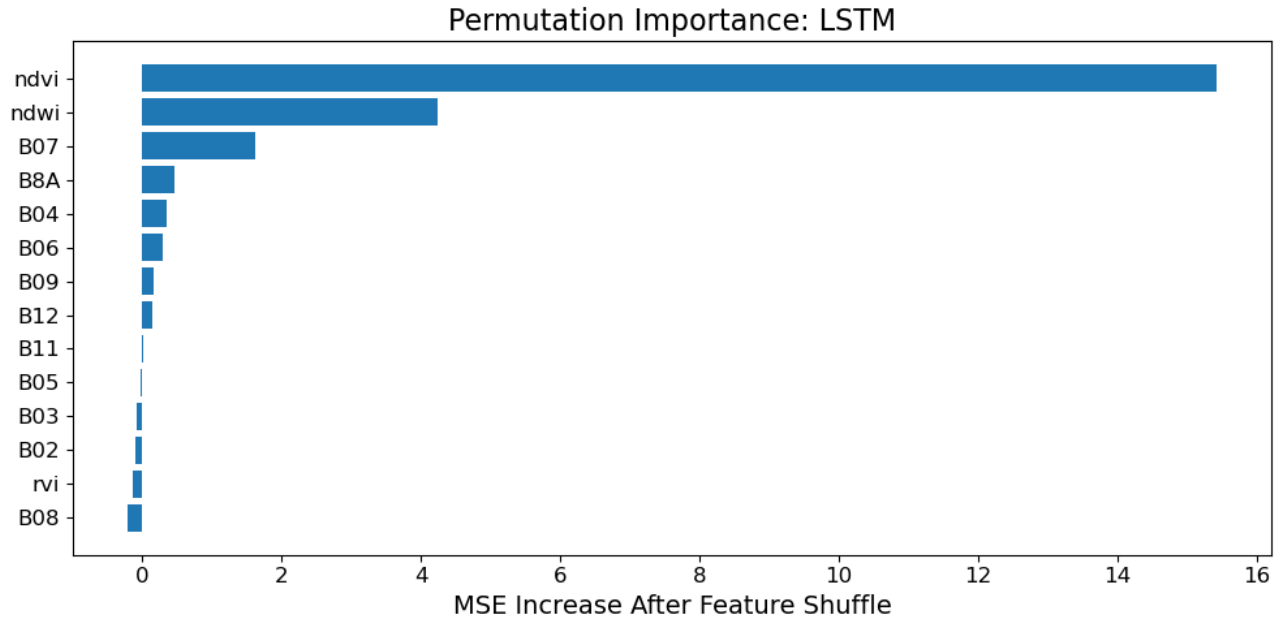

Fig. 12. Permutation Importance for LSTM corn model

The Permutation Importance analysis (Figure 12) shows that NDVI dominates with an MSE increase of approximately 15.5 upon shuffling — far ahead of all other features. NDWI ranks second (~4.2), followed by B07 (~1.8). B8A, B04, and B06 show small but positive contributions (~0.3–0.5), while the remaining features including RVI, B08, B02, B03, B05, B11, and B12 have near-zero importance, indicating the LSTM relies on a narrow set of key inputs.

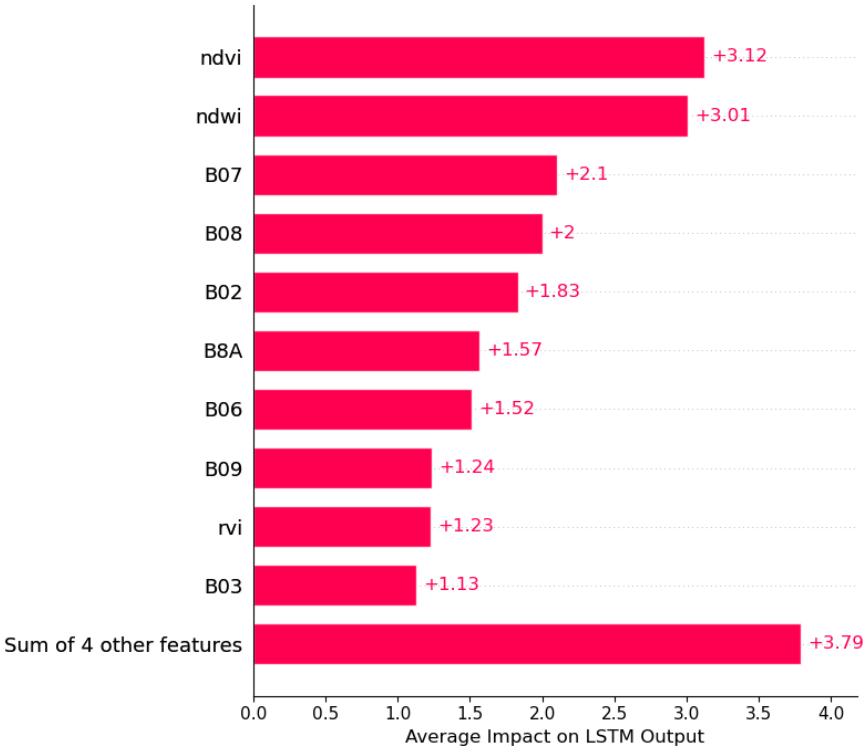

Fig. 13. SHAP explanation for LSTM corn model

The SHAP analysis (Figure 13) paints a notably different picture: while NDVI still leads (+3.12), NDWI is almost equally important (+3.01). More strikingly, the LSTM distributes SHAP impact broadly across many features — B07 (+2.10), B08 (+2.0), B02 (+1.83), B8A (+1.57), B06 (+1.52), B09 (+1.24), RVI (+1.23), and B03 (+1.13) all carry substantial contributions. The "Sum of 4 other features" adds another +3.79. This discrepancy between permutation importance and SHAP suggests that the LSTM captures complex internal interactions among features that are not disrupted by shuffling a single variable, but that nonetheless contribute meaningfully to each individual prediction.

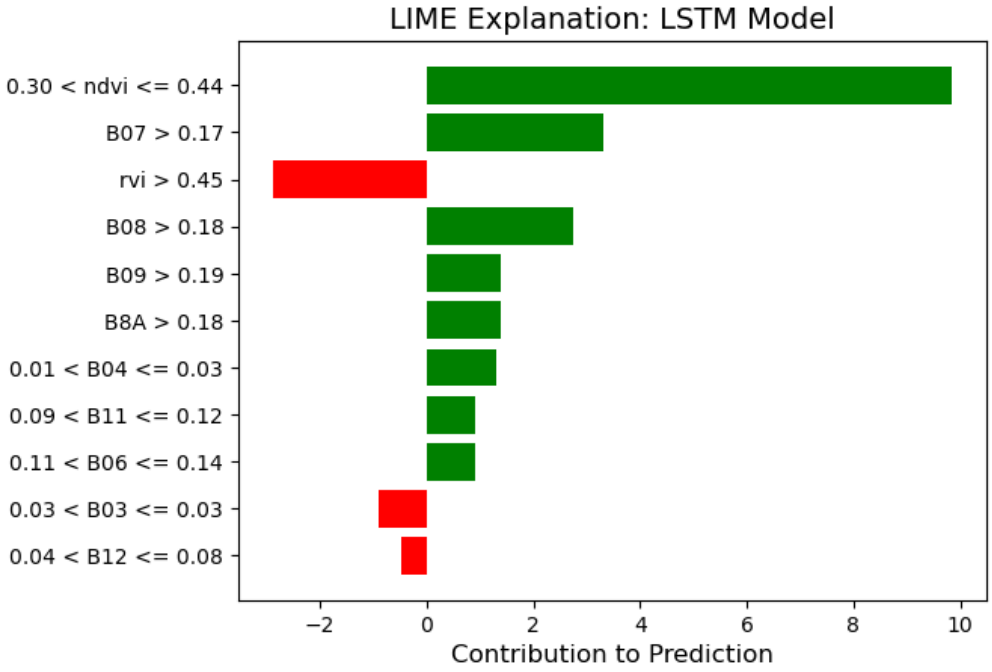

Fig. 14. LIME explanation for LSTM corn model

The LIME explanation (Figure 14) shows that at the local level, NDVI in the range 0.30–0.44 has the strongest positive contribution (~10.0), followed by B07 above 0.17 (~3.5) and B08 above 0.18 (~2.8). Notably, RVI above 0.45 shows a negative contribution (~−2.5, red bar), meaning that for specific corn samples high RVI values actually push the predicted BBCH downward. B09, B8A, B04, B11, B06, and B03 contribute positively but modestly. This negative RVI effect in LIME is consistent with the physical explanation that for row-planted corn, high radar backscatter may reflect exposed soil rather than vegetation density.

3.2.4. Comparison of models for corn

Quantitative evaluation on the test set revealed substantial differences in model performance. The LSTM achieved the lowest mean squared error MSE of 0.2253, significantly outperforming both the feed-forward neural network MSE: 13.3690 and the linear regression baseline MSE: 70.0291. Table 3 summarizes the comparative results for corn models.

Table 3. Comparison of corn models

Model	MSE	RMSE
LSTM	0.2253	0.475
FFNN	13.3690	3.656
Linear Regression	70.0291	8.368

Visual inspection of sample predictions further illustrates these differences. As shown in Figure 15, for the test sample on July 21 the LSTM predicted a BBCH value of 50.1 (actual: 51.0), closely matching the ground truth, while the FFNN and linear regression produced larger deviations of 37.5 and 36.9, respectively.

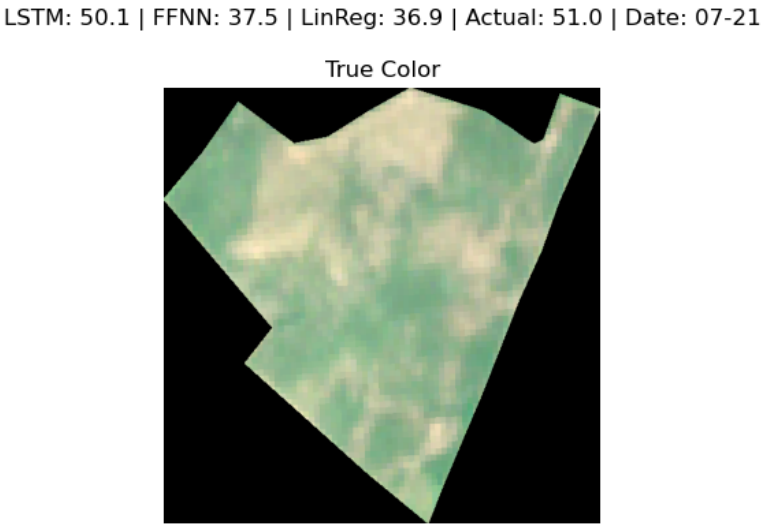

Fig. 15. Predicted and actual BBCH stage on July 21 by corn models

Figure 16 presents an earlier-season example for June 20, where the LSTM prediction of 15.5 (actual: 15.0) again outperformed the alternative models. While the LSTM demonstrated clear superiority in leveraging complex nonlinear patterns, the results of simpler models could be affected by the lack

and diversity of data. As mentioned earlier, the data used did not have measurements for all growth stages, which could affect the results of FFNN and especially Linear Regression.

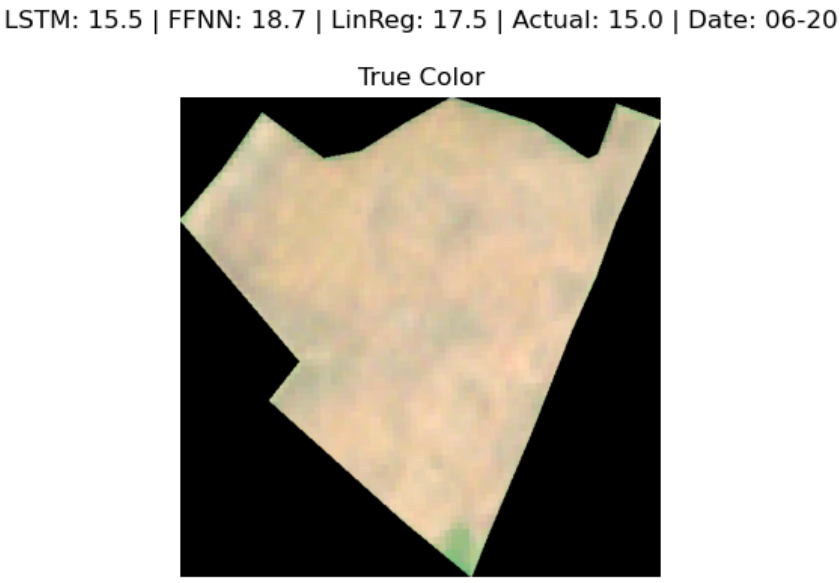

Fig. 16. Predicted and actual BBCH stage on June 20 by corn models

3.2.5. Linear Regression for winter wheat

For winter wheat, the Ridge regression model, a linear regression variant with L2 regularization, was trained on the prepared feature set derived from Sentinel-1/2 data and vegetation indices. In contrast to the corn experiment, model development for winter wheat followed a GroupKFold strategy with K = 6, where the grouping variable was the field identifier. This ensured that observations from the same field never appeared simultaneously in the training and validation partitions, thereby preventing spatial leakage and providing a more realistic estimate of generalization across fields.

Explainable AI methods — permutation importance, SHAP, and LIME — were again applied to quantify feature importance, but this time the results were aggregated across all six folds (mean and standard deviation). As shown in Figures 17–19 for winter wheat, the linear model is driven primarily by Radar-based RVI, as well as by near-infrared and red-edge bands, which are most sensitive to the state of the vegetation canopy. NDVI and NDWI also contribute to the forecasts, although their influence remains more modest compared to RVI.

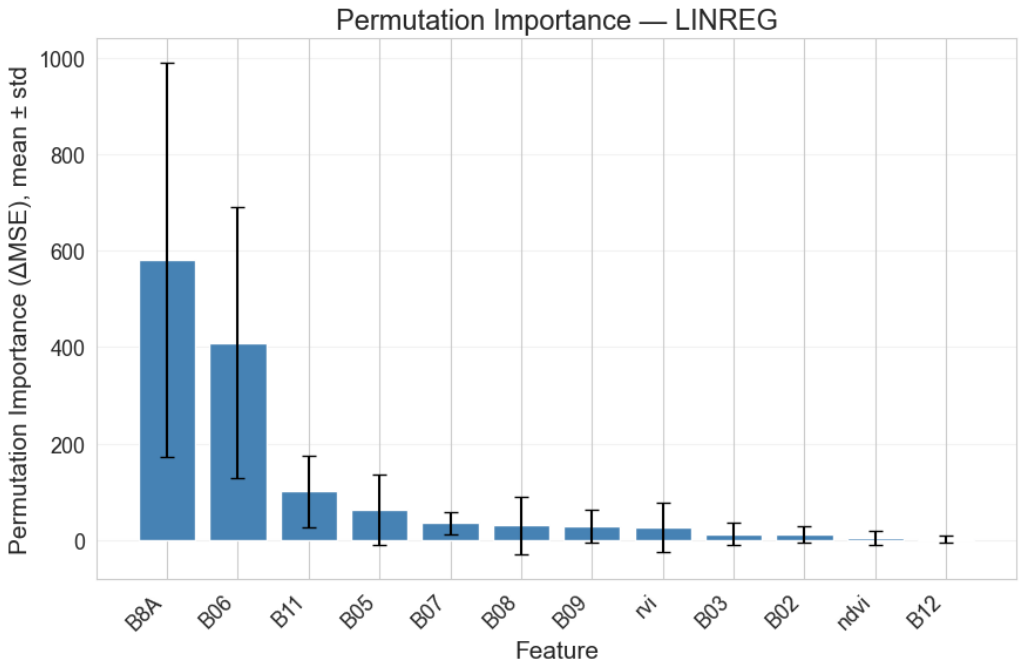

Fig. 17. Permutation Importance for Linear Regression winter wheat model

The Permutation Importance analysis (Figure 17) reveals a fundamentally different feature hierarchy compared to the corn models. The narrow NIR band B8A dominates with a mean ΔMSE of approximately 580, followed by the red-edge band B06 (~400). The SWIR band B11 ranks third (~100), with B05, B07, B08, B09, and RVI forming a mid-tier group (~30–65 each). Notably, NDVI and NDWI — which were the top features for all corn models — show near-zero permutation importance for wheat, and B12 is similarly negligible. The large error bars indicate substantial variability across folds, particularly for B8A and B06.

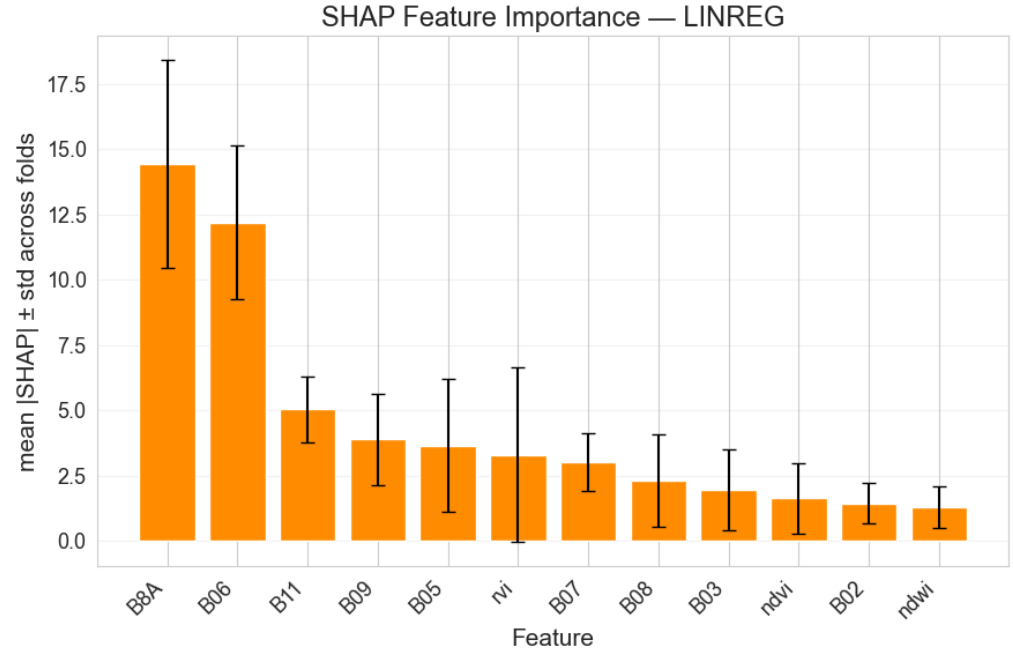

Fig. 18. SHAP explanation for Linear Regression winter wheat model

The SHAP analysis (Figure 18) confirms this pattern: B8A leads with a mean absolute SHAP value of approximately 14.3, followed by B06 (~12.3) and B11 (~5.0). B09, B05, RVI, and B07 form a

secondary group (~3.0–3.9 each), while NDVI (~1.6) and NDWI (~1.2) rank among the least impactful features.

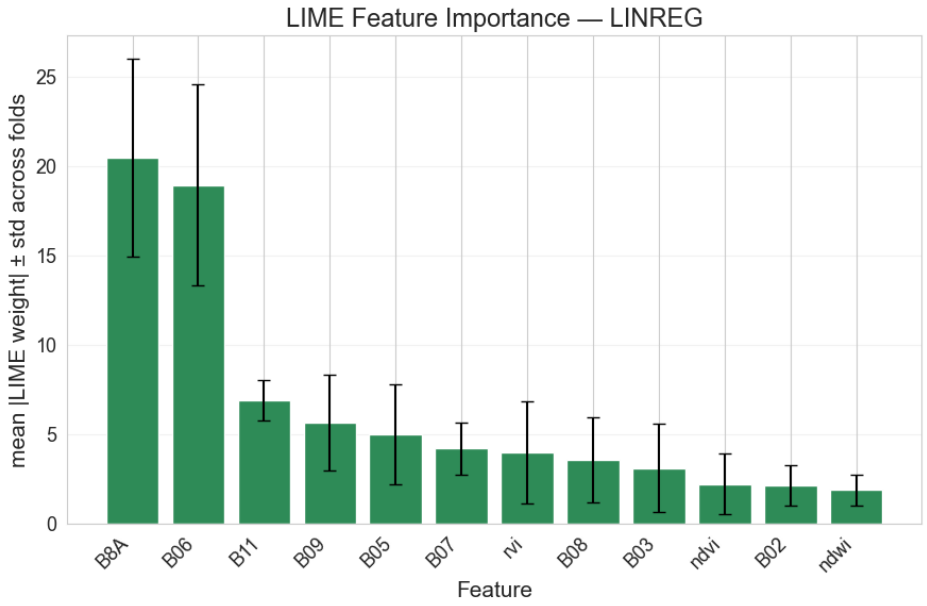

Fig. 19. LIME explanation for Linear Regression winter wheat model

The LIME explanation (Figure 19) provides consistent evidence across folds: B8A (~20.3) and B06 (~18.9) carry the highest mean absolute LIME weights, followed by B11 (~6.9), B09 (~5.5), and B05 (~5.0). RVI (~4.0) shows moderate local influence, while NDVI (~2.1) and NDWI (~1.8) remain at the bottom.

All three XAI methods agree that for winter wheat, the linear model is driven primarily by near-infrared and red-edge bands rather than by the derived vegetation indices that dominated the corn models. This contrast reflects the dense, continuous canopy of winter wheat, which produces strong spectral responses in the NIR and red-edge regions throughout its vegetative development.

3.2.6. Feed-forward neural network for winter wheat

For winter wheat, the feed-forward neural network (FFNN) architecture was optimized using Bayesian hyperparameter search combined with GroupKFold cross-validation (K = 6). The search was performed over the size of the hidden layers and dropout rates, using negative mean squared error as the scoring metric. The best configuration found for the FFNN was:

- first hidden layer size 3980 neurons,
- second hidden layer size 1742 neurons,
- dropout rates 0.076 and 0.5417 after the first and second hidden layers, respectively.

The resulting network topology — input layer of 14 features, two ReLU-activated hidden layers of 3980 and 1742 units with dropout rates of 0.076 and 0.5417 respectively, and a single regression output — was instantiated independently for each of the six cross-validation folds. Per-fold training followed the protocol described in section 3.2.2 (AdamW with lr = 1×10^{-3}, MSE objective, 1000-epoch cap with early stopping on validation MSE). Predictions from the six fold-specific models were averaged at test time to produce the ensemble estimate reported below.

Permutation importance analyses, SHAP, and LIME were calculated for each validation sample and then averaged across all samples. The XAI results in Figures 20–22 show that RVI has a greater

influence than NDVI and NDWI, which differs from the results for corn models. Infrared values also have a significant influence.

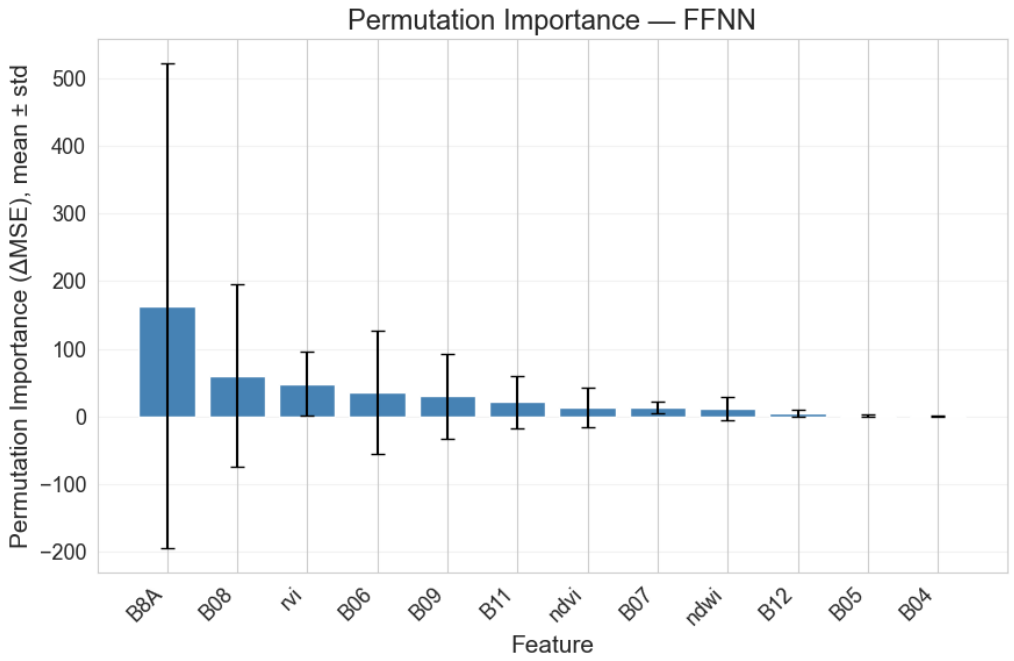

Fig. 20. Permutation Importance for FFNN winter wheat model

The Permutation Importance results (Figure 20) show that B8A again dominates with a mean ΔMSE of approximately 160, though with very large variability across folds (error bars spanning from −200 to +500). B08 ranks second (~58), followed by RVI (~45), B06 (~33), B09 (~28), and B11 (~20). NDVI and NDWI show low permutation importance (~10 each), consistent with the linear model's pattern. B12, B05, and B04 are near zero. The extremely wide error bars — particularly for B8A and B06 — indicate that the FFNN's reliance on these features varies substantially depending on which fields are included in each fold.

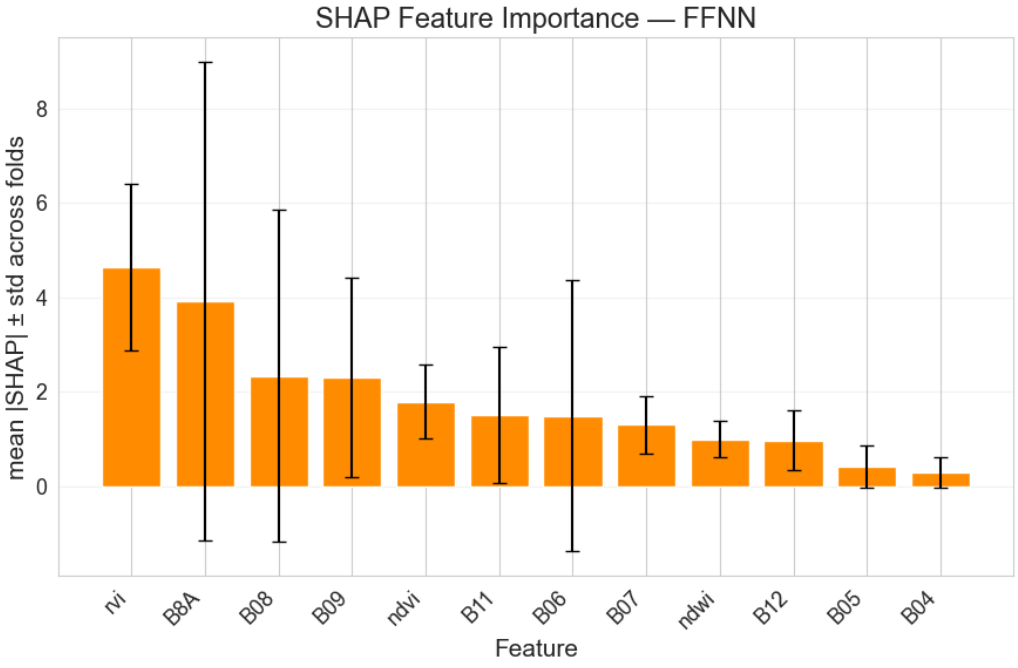

Fig. 21. SHAP explanation for FFNN winter wheat model

The SHAP analysis (Figure 21) reveals a different ranking at the top: RVI leads with the highest mean absolute SHAP value (~4.6), surpassing B8A (~3.9) which dominated in permutation importance. B08 and B09 share third place (~2.3 each), followed by NDVI (~1.7), B11 and B06 (~1.4–1.5), B07 (~1.3), NDWI and B12 (~0.9–1.0), and B05 and B04 at the bottom (~0.3–0.4). The large standard deviations across folds, especially for B8A, indicate high instability in the FFNN's feature utilization.

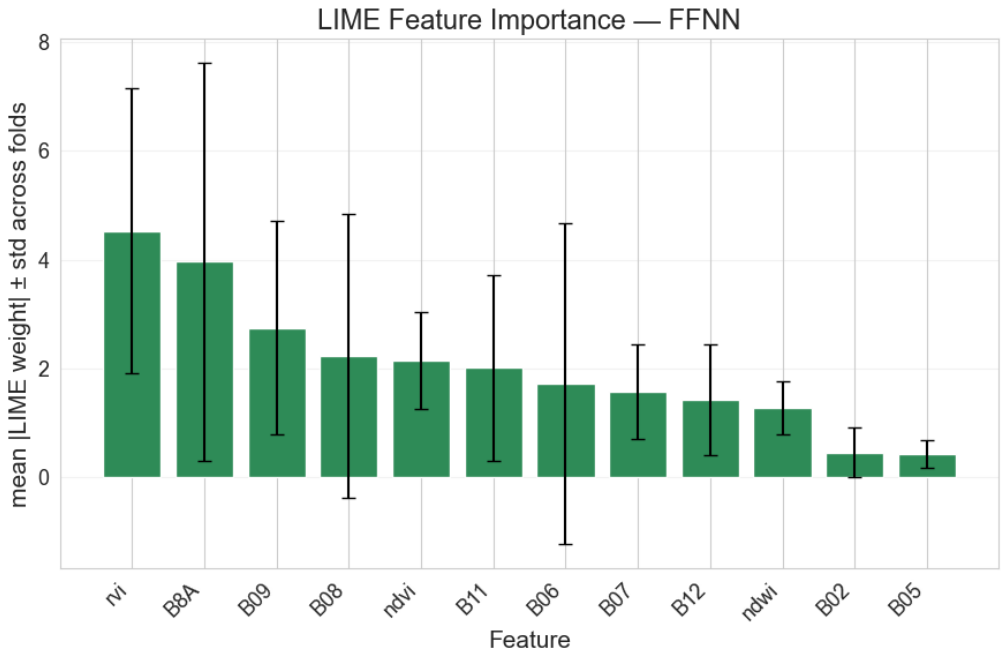

Fig. 22. LIME explanation for FFNN winter wheat model

The LIME analysis (Figure 22) corroborates the SHAP findings, with RVI showing the highest mean absolute LIME weight, confirming that the FFNN for wheat increasingly relies on radar-derived information compared to the linear model. The emergence of RVI as the top SHAP feature — despite ranking only third in permutation importance — suggests that the FFNN uses RVI consistently across individual predictions even though its global impact on MSE is masked by fold-to-fold variability.

3.2.7. Long short-term memory neural network for winter wheat

The LSTM model for winter wheat was also tuned via Bayesian optimization under the same GroupKFold (K = 6) scheme. The best architecture identified by the search was a three-layer LSTM with a hidden state size of 1024 units and a dropout rate of 0 between recurrent layers. On top of the LSTM stack, a fully connected regression head was used, mirroring the design of the corn model (layer normalization, dense layer, ReLU, and a final output neuron).

As in the FFNN case, the LSTM models were trained independently on each of the six folds using AdamW (learning rate 0.001) and MSE loss for up to 1,000 epochs. By processing the input feature vectors as short sequences, the LSTM can capture complex patterns in the winter wheat observations, such as consistent trends of NDVI/NDWI and RVI across successive growth stages.

The XAI methods were again applied iteratively and then aggregated. The results show that LSTM relies heavily on RVI, also NDVI, NDWI and infrared bands also having an impact.

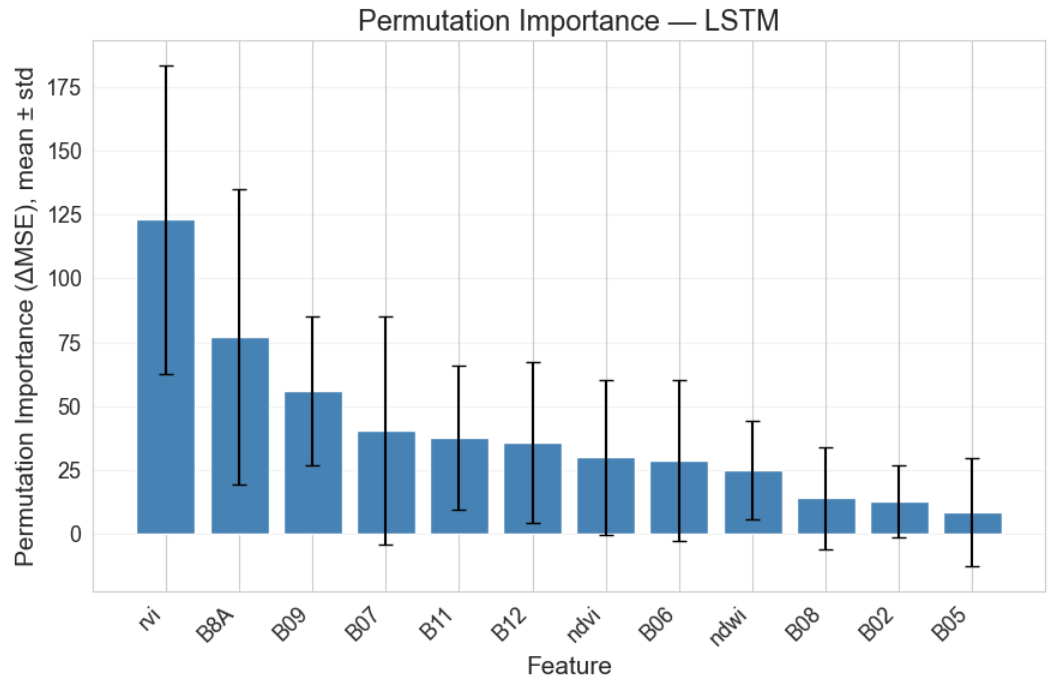

Fig. 23. Permutation Importance for LSTM winter wheat model

The Permutation Importance analysis (Figure 23) shows that RVI is the most influential feature for the wheat LSTM, with a mean ΔMSE of approximately 123 — the first model where RVI surpasses all spectral bands in permutation importance. B8A ranks second (~77), followed by B09 (~55), B07 (~40), B11 (~38), B12 (~35), NDVI (~30), B06 (~29), and NDWI (~25). B08, B02, and B05 contribute least (~7–14). Compared to the FFNN where B8A dominated permutation importance, the LSTM shifts decisively toward RVI, indicating that the gating mechanisms of the LSTM are particularly effective at exploiting radar-derived vegetation structure information.

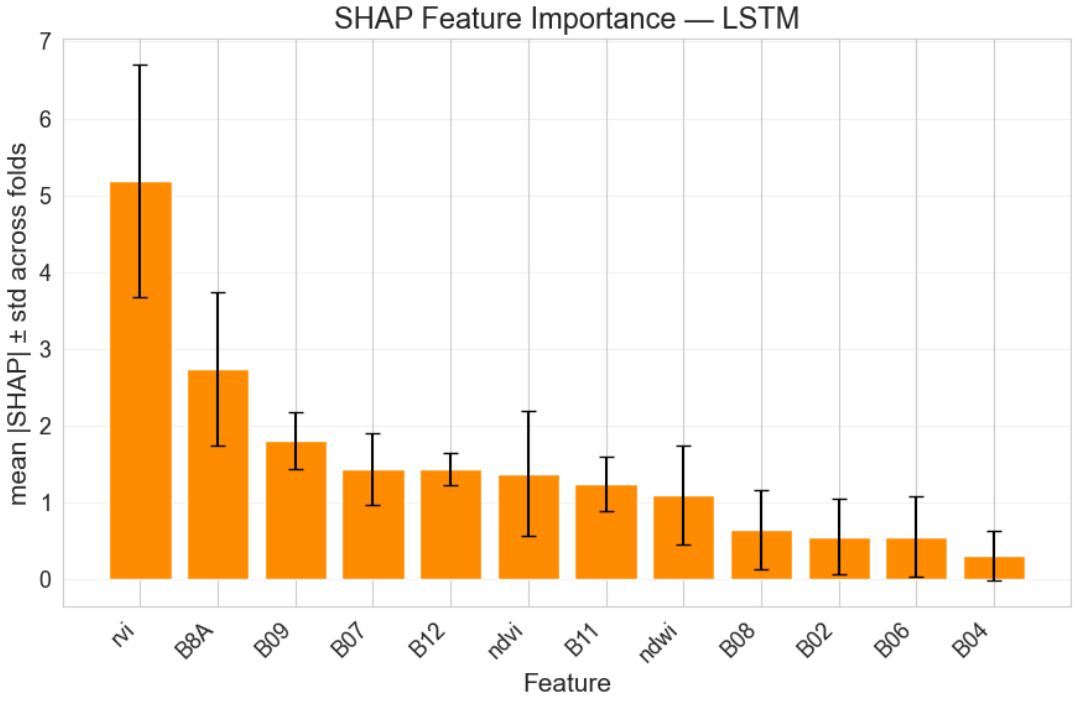

Fig. 24. SHAP explanation for LSTM winter wheat model

The SHAP analysis (Figure 24) confirms RVI's dominance with the highest mean absolute SHAP value (~5.15), nearly double that of B8A (~2.7). B09 (~1.8), B07 and B12 (~1.4 each), NDVI (~1.35), B11 (~1.2), and NDWI (~1.05) form a secondary tier, while B08, B02, B06, and B04 contribute modestly (~0.3–0.6). The gap between RVI and all other features is larger in SHAP than in permutation importance, reinforcing that RVI drives individual LSTM predictions more strongly than any other input.

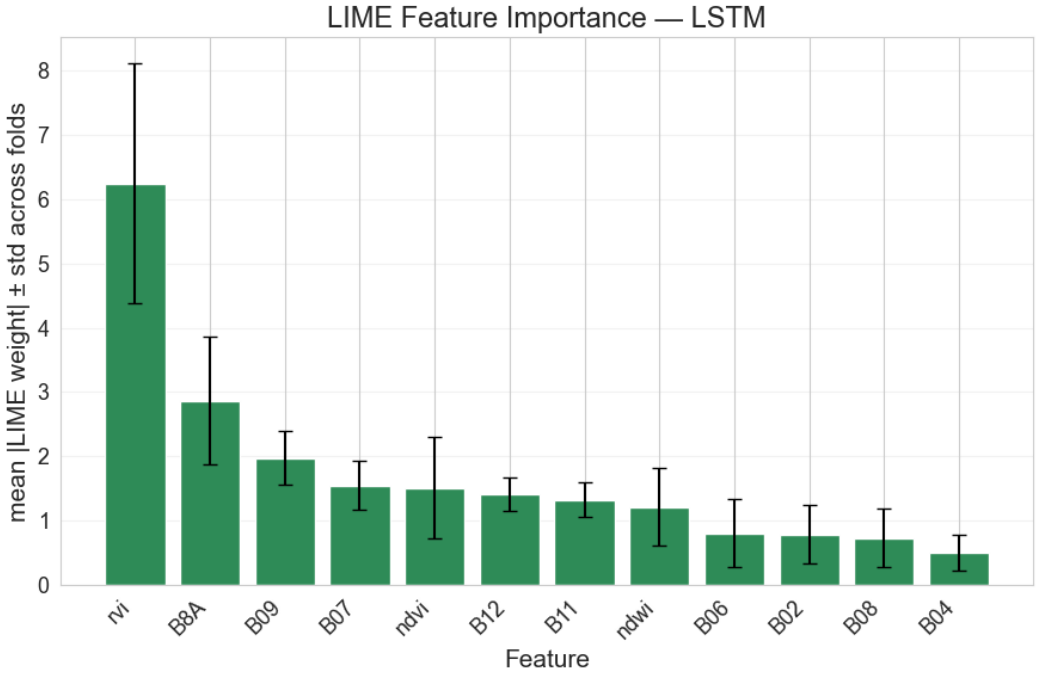

Fig. 25. LIME explanation for LSTM winter wheat model

The LIME analysis (Figure 25) provides consistent confirmation: RVI leads with the highest mean absolute LIME weight (~6.2), followed by B8A (~2.8), B09 (~2.0), B07 (~1.5), NDVI (~1.5), B12 (~1.4), B11 (~1.3), and NDWI (~1.2). B06, B02, B08, and B04 remain at the bottom (~0.5–0.8).
All three XAI methods agree unanimously that for the wheat LSTM, RVI is the single most important predictor — a sharp contrast to the corn LSTM where NDVI dominated. This cross-crop divergence is physically consistent: the dense continuous canopy of winter wheat produces strong and consistent radar backscatter patterns that the LSTM can effectively model, whereas the row-planted corn canopy leaves soil exposed, making optical indices more informative.

3.2.8. Comparison of models for winter wheat

The performance of the three models for winter wheat was evaluated using GroupKFold cross-validation with field-wise separation of samples. Table 4 summarizes the cross-validated mean squared error (MSE) obtained for each model across all folds.

Table 4. Winter wheat models comparison by folds

Fold \ AI Model	LSTM	FFNN	Linear Regression
1	71.9302	102.8488	97.1688
2	78.5975	116.1939	133.9850
3	52.7736	102.0216	251.8527
4	22.6941	34.3596	123.2637
5	115.2274	174.6012	162.1918
6	59.4903	79.5949	120.5119
Mean MSE and standard deviation	**66.7855 ± 30.6797**	**101.6033 ± 45.9334**	**148.1623 ± 55.0170**

The LSTM model demonstrated the best overall performance, achieving the lowest average validation error with a mean MSE of 66.78 and a standard deviation of 30.68. Despite some variability between folds, the LSTM consistently outperformed the other approaches, confirming its superior ability to capture the dynamics of vegetation development.

The FFNN showed moderate performance with a mean MSE of 101.60 ± 45.93. While the FFNN improved upon the linear baseline, its results were less stable across folds and exhibited larger variance, indicating sensitivity to field-to-field variability.

The Linear regression model produced the weakest results, with a mean MSE of 148.16 ± 55.02, confirming that linear relationships alone are insufficient to model the complex spectral–phenological dependencies of crop growth.

To further assess the practical behavior of the models, predictions were computed by averaging the outputs of models trained on all folds. Three representative single-date examples and an aggregated comparison plot are presented in Figures 26–29, where predicted BBCH stages are compared to ground-truth observations and accompanied by true-color Sentinel-2 imagery. Figure 26 illustrates an early-season case on 30 May, where the true BBCH value was 37.0 and the LSTM ensemble produced the most accurate prediction, while the FFNN and linear regression overestimated the growth stage.

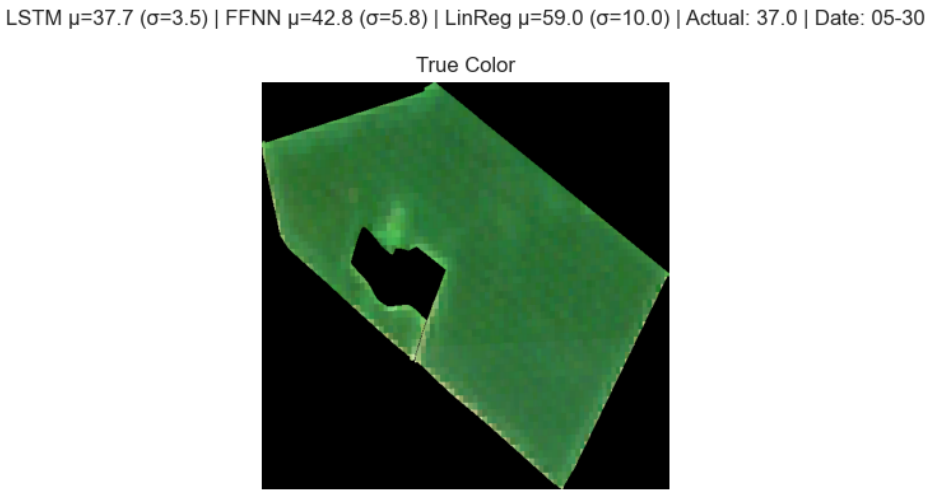

Fig. 26. Predicted and actual BBCH stage on May 30 by winter wheat models

Figure 27 presents a mid-season example on 12 June, with a ground-truth value of BBCH 55.0. At this intermediate phenological phase, all three models converge toward the observed value, and even the linear regression produces a near-accurate prediction.

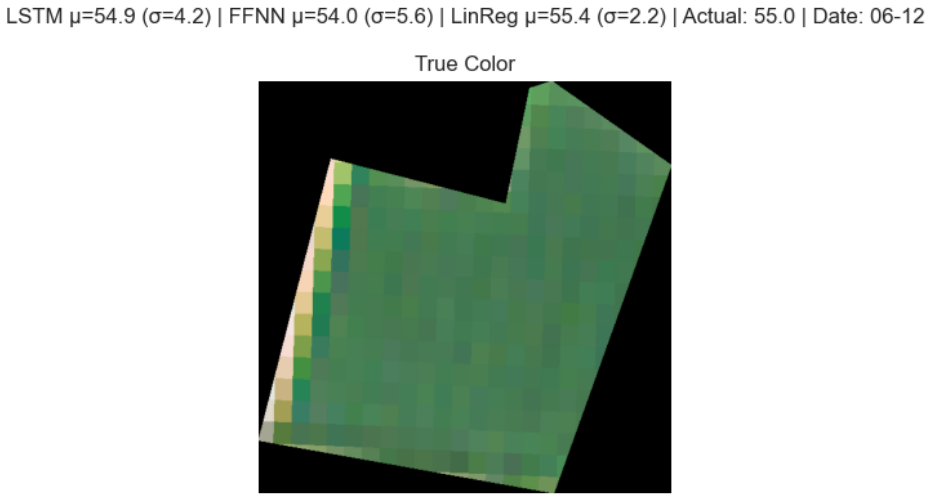

Fig. 27. Predicted and actual BBCH stage on June 12 by winter wheat models

Figure 28 shows a late-season case on 25 June, where the true BBCH value reached 69.0. In this case, all three models systematically underestimate the growth stage, reflecting the limited representation of late-season observations in the training data.

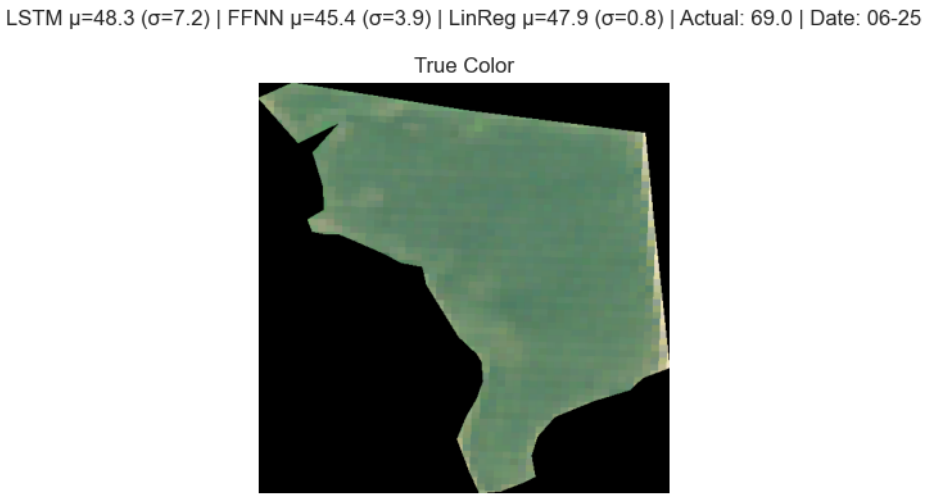

Fig. 28. Predicted and actual BBCH stage on June 25 by winter wheat models

Figure 29 aggregates all per-sample predictions on a single comparison plot, providing a global view of model accuracy across the full phenological range and highlighting the systematic divergence between the LSTM, the FFNN, and the linear regression at the upper end of the BBCH scale.

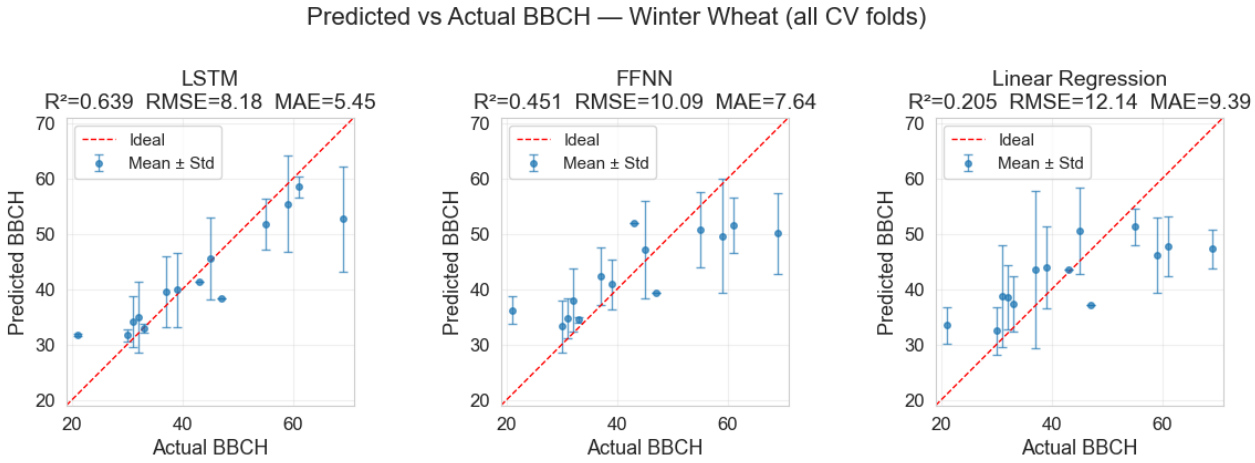

Fig. 29. Predicted and actual BBCH stage comparison for winter wheat

As illustrated in Figures 26–29, the models demonstrate varying levels of accuracy depending on the phenological phase. In the first example (30 May, BBCH 37.0), the LSTM ensemble prediction ($\mu = 37.7$, $\sigma = 3.5$) closely matches the observed value, while the FFNN ($\mu = 42.8$, $\sigma = 5.8$) and the linear regression model ($\mu = 59.0$, $\sigma = 10.0$) overestimate the growth stage considerably. In the mid-season example (12 June, BBCH 55.0), all three models converge toward the ground truth, with the LSTM ($\mu = 54.9$) and FFNN ($\mu = 54.0$) producing near-accurate predictions and even the linear regression ($\mu = 55.4$) performing well at this intermediate stage. However, in the late-season case (25 June, BBCH 69.0), all models systematically underestimate the true growth stage, with the LSTM predicting $\mu = 48.3$, the FFNN $\mu = 45.4$, and linear regression $\mu = 47.9$. This progressive degradation in accuracy at higher BBCH values can be attributed to the uneven temporal distribution of ground-truth observations in the training set: while the early-to-mid growth stages (BBCH 30–55) are well represented, samples corresponding to late-season stages (BBCH $\geq$ 60) are significantly underrepresented. As a result, the models lack sufficient examples to learn the spectral-phenological

relationships characteristic of advanced maturation phases, leading to a systematic bias toward lower predicted values.

Overall, these results confirm that the LSTM provides the most accurate and robust predictions for winter wheat growth stages, particularly during active growth phases. The FFNN offers moderate accuracy and shows higher uncertainty, while the linear regression model remains the least reliable.

3.2.9. Comparison of models trained using selected parameters for winter wheat

Following the analysis of feature importance using permutation importance, SHAP, and LIME across all folds of the full-feature models, a subset of the most impactful features was identified for winter wheat. The selected features included six Sentinel-2 spectral bands — B06, B07, B8A, B09, B11, and B12 - along with three derived vegetation indices: NDVI, NDWI, and RVI. This reduced the input dimensionality from 14 features to 9.

All three models were retrained from scratch using the same GroupKFold (K=6) cross-validation strategy, Bayesian hyperparameter optimization procedure, and training protocols as in the full-feature experiments. Table 5 summarizes the comparative results.

Table 5. Comparison of winter wheat model performance using different feature sets

Model	**Features**	**MSE (mean ± std)**	**RMSE**	**MAE**	**R^2**
LSTM	All 14	66.79 ± 30.68	8.18	5.45	0.639
LSTM	Selected 9	72.11 ± 48.23	8.49	6.29	0.598
FFNN	All 14	101.60 ± 45.93	10.09	7.64	0.451
FFNN	Selected 9	86.40 ± 40.58	9.30	7.56	0.520
Linear Regression	All 14	148.16 ± 55.02	12.14	9.39	0.205
Linear Regression	Selected 9	142.12 ± 79.83	11.92	9.24	0.216

The results reveal a nuanced picture. The LSTM model experienced a slight degradation in performance when trained on the reduced feature set, with the mean MSE increasing from 66.79 to 72.11 and R^2 decreasing from 0.639 to 0.598. This suggests that the LSTM benefits from the full spectral information and is capable of effectively leveraging even marginally informative features through its gating mechanisms. Notably, the standard deviation of MSE across folds nearly doubled (from 30.68 to 48.23), indicating less stable performance with the reduced feature set.

In contrast, the FFNN showed a clear improvement with feature selection, with the mean MSE decreasing from 101.60 to 86.40 and R^2 increasing from 0.451 to 0.520. This improvement suggests that the FFNN was previously overfitting to noisy or redundant features in the full 14-dimensional input space, and the removal of less informative bands allowed the network to focus on the most discriminative spectral signals. The fold-to-fold variability also decreased (std from 45.93 to 40.58), indicating more consistent generalization.

The Linear Regression model showed marginal improvement with selected features, with MSE decreasing slightly from 148.16 to 142.12 and R^2 improving from 0.205 to 0.216. However, the increase in standard deviation (from 55.02 to 79.83) indicates higher variability across spatial partitions, suggesting that the linear model remains fundamentally limited in capturing the complex nonlinear relationships between spectral features and BBCH stages regardless of feature selection.

These findings indicate that feature selection has a model-dependent effect: architectures prone to overfitting (such as FFNN with limited training data) benefit from dimensionality reduction, while more expressive models like LSTM can extract useful information from the full feature set. For practical deployment with limited ground-truth data, a reduced feature set combined with an FFNN architecture may offer a favorable trade-off between model complexity and prediction accuracy.

3.2.10. Results analysis

Experimental results obtained for corn and winter wheat clearly demonstrate the advantages of deep learning models, particularly sequence-based architectures, over traditional regression approaches for predicting crop growth stages from satellite data.

For corn, evaluation on an independent test set revealed a significant advantage for the LSTM model, which achieved an extremely low MSE of 0.2253, compared to 13.3690 for FFNN and 70.0291 for the linear regression model. Visual analysis of the predictions (Figures 15–16) further confirmed that the LSTM accurately captures the true BBCH dynamics, while FFNN and, especially, linear regression exhibit significant deviations. These results indicate that the gating mechanisms of the LSTM architecture are essential for modelling corn phenology accurately even with limited training data.

However, the reliability of the simpler models for corn was seriously undermined by the limited size and uneven distribution of the dataset across growth stages. With only 14 total observations across 6 fields and coverage limited to BBCH stages 10–51, FFNN and linear regression were unable to generalize across the entire phenological cycle. In contrast, the LSTM was able to partially compensate for this limitation by exploiting complex feature interactions through its internal cell state.

For winter wheat, a more robust experimental protocol based on GroupKFold cross-validation (K=6) with field-level separation was applied. The LSTM again performed best, achieving a cross-validated mean MSE of 66.79 ± 30.68 (RMSE = 8.18, MAE = 5.45, $R^2 = 0.639$), outperforming both the FFNN (MSE = 101.60 ± 45.93, RMSE = 10.09, MAE = 7.64, $R^2 = 0.451$) and linear regression (MSE = 148.16 ± 55.02, RMSE = 12.14, MAE = 9.39, $R^2 = 0.205$). The relatively large standard deviations across folds indicate significant field-to-field variability, which is expected given the diversity of environmental conditions and management practices across spatial locations.

An additional experiment was conducted to assess whether reducing the input feature space to only the most influential variables could improve model performance. Based on XAI analysis, a subset of 9 features (B06, B07, B8A, B09, B11, B12, NDVI, NDWI, RVI) was selected. The FFNN showed a clear improvement with selected features (MSE decreased from 101.60 to 86.40, R^2 increased from 0.451 to 0.520), suggesting that the full-feature FFNN was overfitting to noisy inputs. In contrast, the LSTM experienced a slight degradation (MSE increased from 66.79 to 72.11), indicating that it benefits from the full spectral information. Linear regression showed marginal improvement (MSE

from 148.16 to 142.12). These findings demonstrate that feature selection effectively reduced noise for the shallower architecture. LSTM and linear regression showed only marginal changes, suggesting that the discarded bands carry complementary information for the larger model while contributing little for the linear one.

Analysis of winter wheat samples (Figures 26–29) further reveals the strengths and weaknesses of each model at different growth stages. At mid-season stages (e.g., BBCH 37 and BBCH 55), the LSTM and FFNN provide accurate predictions with relatively low uncertainty, while the linear regression model exhibits a stronger bias. However, at late stages (BBCH 69), all three models systematically underestimate the true growth stage. This progressive degradation in accuracy at higher BBCH values can be attributed to the uneven temporal distribution of ground-truth observations: while the early-to-mid growth stages (BBCH 30–55) are well represented, samples corresponding to late-season stages (BBCH ≥ 60) are significantly underrepresented.

Table 6. Comparison of results with other published studies

Study	Objective	Crop	Data	Best Method	Metric	Result
Newete et al. [22]	Phenology-based crop classification	Wheat	S-2	RF + K-means	OA / PA	83.58% OA; 92.52% PA (heading)
Jiao et al. [25]	Crop condition tracking via SAR+optical fusion	Canola	S-1 + S-2	CSDM (canopy structure dynamics model)	R^2	R^2 = 0.88 (biomass, early season)
Tian et al. [23]	Wheat yield estimation with remote sensing + meteo	Wheat	S-2 + meteo	LSTM	—	LSTM outperformed regression; time-step critical
Yang et al. [26]	Wheat yield estimation using multispectral + hyperspectral	Wheat	S-2 + ZY-1	LSTM	RMSE (t/ha)	LSTM: 0.201;
Joshi et al. [24]	Explainable wheat yield prediction	Wheat	S-2 + weather	Bi-LSTM + SHAP	R^2 / RMSE	R^2=0.81; RMSE=0.24 t/ha; Bi-LSTM>LSTM>RF
Dhillon et al. [27]	Yield prediction combining RF with crop models	Wheat, OSR	NDVI + climate	RF + crop model	R^2	NDVI: R^2<0.65; NDVI+climate: R^2>0.70
This work	**BBCH growth stage regression from satellite data**	**Corn**	**S-1 + S-2**	**LSTM**	**MSE**	**MSE=0.23;**
This work	**BBCH growth stage regression from satellite data**	**Wheat**	**S-1 + S-2**	**LSTM**	**R^2/RMSE/ MAE/MSE**	**R^2=0.639; RMSE=8.18; MAE=5.45, MSE = 66.79 ± 30.68**

The comparison presented in Table 6 reveals several important patterns. First, across all reviewed studies, the hierarchy of model performance consistently shows that LSTM and other deep learning

architectures outperform traditional machine learning methods and linear regression. Yang, Chao et al. [26] found LSTM superior to RF, GBDT, and SVR for wheat yield estimation, while Joshi, Abhasha et al. [24] reported that Bi-LSTM outperformed standard LSTM and Random Forest. The present work confirms this trend for both corn and winter wheat across all evaluation metrics.

Second, studies utilizing combined Sentinel-1 and Sentinel-2 data consistently achieve better results than those relying on a single data source. Jiao et al. [25] demonstrated that the SAR-derived vegetation condition was strongly correlated with biomass ($R^2 = 0.88$) when optical and radar data were combined. The present work similarly uses fused radar and optical features, with XAI analysis revealing that the relative importance of RVI (radar) versus NDVI/NDWI (optical) varies by crop type: optical indices dominate for corn due to the row-planting system exposing soil, while RVI becomes more influential for the dense canopy of winter wheat.

Third, Dhillon, Maninder Singh et al. [27] demonstrated that NDVI alone was insufficient for accurate crop predictions ($R^2 < 0.65$), and that combining vegetation indices with additional data sources improves performance. This finding is consistent with the present work, where the full 14-feature model (combining optical bands, radar bands, and derived indices) generally outperformed the reduced 9-feature model for the LSTM, although the FFNN benefited from feature selection due to its higher susceptibility to overfitting with limited training data.

An important distinction between this work and the reviewed studies is the prediction target itself. While Newete et al. [22] performed classification of phenological stages as discrete categories, and other studies focused on yield estimation (t/ha), this work directly predicts the BBCH growth stage as a continuous numeric value. This regression-based formulation provides more granular information about crop development status than categorical classification and addresses a different practical need than yield forecasting.

A notable difference between the two crops is the relative importance of optical vegetation indices (NDVI, NDWI) and the radar vegetation index (RVI). For corn, XAI analyses consistently showed that NDVI and NDWI dominate model predictions, while RVI has only a minor impact. This can be explained by the row-planted corn system, which leaves significant areas of exposed soil visible between rows, especially during early and mid-growth stages. Given the 10-meter resolution of Sentinel-1 imagery, RVI values often represent a mixture of vegetation and soil reflectance. For winter wheat, the situation is fundamentally different: the dense, continuous canopy covers the soil surface early in the growth phase, making radar backscatter highly sensitive to canopy structure and biomass. This explains the strong dependence of FFNN and LSTM models on RVI for winter wheat, especially during the booting and heading stages.

Taken together, the experiments confirm the LSTM as the strongest performer across both crops, owing to its capacity for representing intricate non-linear couplings between inputs and target. The FFNN occupies a middle ground and clearly gains from a reduced feature set when ground-truth samples are scarce. Ridge regression, by contrast, proves the least dependable choice — its linear formulation simply cannot accommodate the entangled spectral-phenological dynamics that govern crop development.

Conclusions

1. Literature review and selection of data sources. A systematic review of remote-sensing and machine-learning approaches to crop growth stage estimation identified Sentinel-1 (C-band SAR) and Sentinel-2 (multispectral optical) as the most suitable, freely available, and complementary data sources for this work, providing radar backscatter under any weather conditions and reflectance in the VIS, NIR, and SWIR ranges required for vegetation index computation.

2. Data collection and preprocessing pipeline. A reproducible pipeline was implemented that integrates Smart Agrometer ground truth (crop type, BBCH stage, GPS coordinates, timestamp) with satellite observations acquired through the Copernicus Data Space API using a ±2-day temporal tolerance, spatially averages spectral and radar values over field polygons, computes NDVI, NDWI, and RVI, and normalises the resulting 14-feature vector for model input.

3. Model architectures and performance. Three architectures of increasing complexity (Ridge regression, FFNN, LSTM) were implemented and trained. For corn, the LSTM achieved MSE = 0.23 on an independent test set, against 13.37 (FFNN) and 70.03 (Ridge). For winter wheat (6-fold GroupKFold with field-level separation), the LSTM achieved $R^2 = 0.639$, RMSE = 8.18, MAE = 5.45, outperforming the FFNN ($R^2 = 0.451$) and Ridge ($R^2 = 0.205$).

4. Evaluation protocol. Cross-validation with strict field-level separation was confirmed to be essential: it prevents spatial data leakage between train and validation sets, and the standard deviations observed across folds (MSE = 66.79 ± 30.68 for the LSTM on wheat) quantify the realistic field-to-field variability that any deployment of such models would encounter.

5. XAI analysis and feature selection. Permutation Importance, SHAP, and LIME consistently showed crop-specific feature hierarchies: NDVI and NDWI dominate for corn (whose row planting exposes soil and weakens the radar signal), while RVI together with NIR (B8A) and red-edge (B06, B07) bands dominate for winter wheat (whose continuous canopy maximises SAR sensitivity). The XAI-driven feature selection (from 14 to 9 features) improved FFNN performance (R^2 from 0.451 to 0.520) but slightly degraded the LSTM, indicating that the benefit of dimensionality reduction is model-dependent.

6. Limitations and future work. The main limitation of this study is the small size and uneven temporal coverage of the ground-truth dataset, in particular the underrepresentation of late-season stages (BBCH ≥ 60), which caused systematic underestimation at advanced maturation phases. Extending the dataset to cover heading, flowering, and ripening stages, as well as additional growing seasons and field sites, is expected to substantially improve model accuracy and generalisation.

List of references

1. *Monitoring Corn Nitrogen Concentration from Radar (C-SAR), Optical, and Sensor Satellite Data Fusion.* Lapaz, Adrián Olveira, et al. Remote Sensing. 2023, 15(3), 824. Available from: doi: http://dx.doi.org/10.3390/rs15030824
2. *Advancements in Utilizing Image-Analysis Technology for Crop-Yield Estimation.* Yu, Feng, et al. Remote Sensing. 2024, 16(6), 1003. Available from: doi: http://dx.doi.org/10.3390/rs16061003
3. *A Meta-Analysis of Remote Sensing Technologies and Methodologies for Crop Characterization.* Bahrami, Hazhir, McNairn, Heather and Homayouni, Saeid. Remote Sensing. 2022, 14(22), 5633. Available from: doi: http://dx.doi.org/10.3390/rs14225633
4. *Artificial Intelligence Techniques in Crop Yield Estimation Based on Sentinel-2 Data: A Comprehensive Survey.* Fatih, Muhammet Aslan, Sabanci, Kadir and Aslan, Busra. Sustainability. 2024, 16(18), 8277. Available from: doi: http://dx.doi.org/10.3390/su16188277
5. *Estimating crop primary productivity with Sentinel-2 and Landsat 8 using machine learning methods trained with radiative transfer simulations.* Wolanin, Aleksandra, et al. Remote Sensing of Environment. 2019, 225, 441-457. Available from: doi: http://dx.doi.org/10.1016/j.rse.2019.03.002
6. *Predicting Crop Yield with Machine Learning: An Extensive Analysis of Input Modalities and Models on a Field and Sub-Field Level.* Pathak, Deepak, et al. IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium. 2023, 2767–2770. Available from: doi: http://dx.doi.org/10.48550/arXiv.2308.08948
7. *Scalable Prediction of Crop BBCH Stages Using Sentinel-2 Plant Phenology Index and Machine Learning.* Neubauer, Thomas, Koelbing, Marlene and Manschadi, Ahmad. SSRN Electronic Journal. 2025. Available from: doi: http://dx.doi.org/10.2139/ssrn.5360947
8. *Evaluation of Sentinel-1 and 2 time series for predicting wheat and rapeseed phenological stages.* Mercier, Audrey, et al. ISPRS Journal of Photogrammetry and Remote Sensing. 2020, 163, 231-256. Available from: doi: http://dx.doi.org/10.1016/j.isprsjprs.2020.03.009
9. *Status of Phenological Research Using Sentinel-2 Data: A Review.* Misra, Gourav, Cawkwell, Fiona and Wingler, Astrid. Remote Sensing. 2020, 12(17), 2760. Available from: doi: http://dx.doi.org/10.3390/rs12172760
10. *Exploring the effects of crop growth differences on radar vegetation index response and crop height estimation using dynamic monitoring model.* Wang, Bo, et al. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2024, X-1-2024, 225-232. Available from: doi: http://dx.doi.org/10.5194/isprs-annals-X-1-2024-225-2024
11. *Crop Monitoring Using Satellite/UAV Data Fusion and Machine Learning.* Maimaitijiang, Maitiniyazi, et al. Remote Sensing. 2020, 12(9), 1357. Available from: doi: http://dx.doi.org/10.3390/rs12091357
12. *Toward Precision in Crop Yield Estimation Using Remote Sensing and Optimization Techniques.* Awad, Mohamad M. Agriculture. 2019, 9(3), 54. Available from: doi: http://dx.doi.org/10.3390/agriculture9030054
13. *USA Crop Yield Estimation with MODIS NDVI: Are Remotely Sensed Models Better than Simple Trend Analyses?* Johnson, David M., et al. Remote Sensing. 2021, 13(21), 4227. Available from: doi: http://dx.doi.org/10.3390/rs13214227

14. *Synergistic integration of optical and microwave satellite data for crop yield estimation.* Mateo-Sanchis, Anna, et al. Remote Sensing of Environment. 2019, 234, 111460. Available from: doi: http://dx.doi.org/10.1016/j.rse.2019.111460
15. Sabini, Mark, Rusak, Gili and Ross, Brad. *Understanding Satellite-Imagery-Based Crop Yield Predictions.* Stanford University, CS231n. 2017. [Online] [Cited: 22 February 2026.] Available from: https://marksabini.com/files/cs231n__Understanding_Satellite-Imagery-Based_Crop_Yield_Predictions__report.pdf
16. *Vegetation descriptors from Sentinel-1 SAR data for crop growth monitoring.* Bao, Xin, et al. ISPRS Journal of Photogrammetry and Remote Sensing. 2023, 203, 86-114. Available from: doi: http://dx.doi.org/10.1016/j.isprsjprs.2023.07.023
17. *Sentinel-1 interferometric coherence and backscattering analysis for crop monitoring.* Nasirzadehdizaji, Rouhollah, et al. Computers and Electronics in Agriculture. 2021, 185, 106118. Available from: doi: http://dx.doi.org/10.1016/j.compag.2021.106118
18. *Calibrating vegetation phenology from Sentinel-2 using eddy covariance, PhenoCam, and PEP725 networks across Europe.* Tian, Feng, et al. Remote Sensing of Environment. 2021, 260, 112456. Available from: doi: http://dx.doi.org/10.1016/j.rse.2021.112456
19. *Dual polarimetric radar vegetation index for crop growth monitoring using Sentinel-1 SAR data.* Mandal, Dipankar, et al. Remote Sensing of Environment. 2020, 247, 111954. Available from: doi: http://dx.doi.org/10.1016/j.rse.2020.111954
20. *Corn Phenology Detection Using the Derivative Dynamic Time Warping Method and Sentinel-2 Time Series.* Ye, Junyan, et al. Remote Sensing. 2023, 15(14), 3456. Available from: doi: http://dx.doi.org/10.3390/rs15143456
21. *Near real-time detection and forecasting of within-field phenology of winter wheat and corn using Sentinel-2 time-series data.* Liao, Chunhua, et al. ISPRS Journal of Photogrammetry and Remote Sensing. 2023, 196, 105-119. Available from: doi: http://dx.doi.org/10.1016/j.isprsjprs.2022.12.025
22. *Phenology-based winter wheat classification for crop growth monitoring using multi-temporal Sentinel-2 satellite data.* Newete, Solomon W., et al. The Egyptian Journal of Remote Sensing and Space Sciences. 2024, 27(4), 695-704. Available from: doi: http://dx.doi.org/10.1016/j.ejrs.2024.10.001
23. *An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, PR China.* Tian, Haifeng, et al. Agricultural and Forest Meteorology. 2021, 310, 108629. Available from: doi: http://dx.doi.org/10.1016/j.agrformet.2021.108629
24. *An explainable Bi-LSTM model for winter wheat yield prediction.* Joshi, Abhasha, et al. Frontiers in Plant Science. 2025, 15, 1491493. Available from: doi: http://dx.doi.org/10.3389/fpls.2024.1491493
25. *Integrating Sentinel-1 SAR and Sentinel-2 optical imagery with a crop structure dynamics model to track crop condition.* Jiao, Xianfeng, et al. International Journal of Remote Sensing. 2022, 43(17), 6509-6537. Available from: doi: http://dx.doi.org/10.1080/01431161.2022.2142077
26. *Wheat yield estimation using remote sensing data based on machine learning approaches.* Yang, Chao, et al. Frontiers in Plant Science. 2022, 13, 1090970. Available from: doi: http://dx.doi.org/10.3389/fpls.2022.1090970

27. *Integrating random forest and crop modeling improves the crop yield prediction of winter wheat and oil seed rape.* Dhillon, Maninder Singh, et al. Frontiers in Remote Sensing. 2023, 3, 1010978. Available from: doi: http://dx.doi.org/10.3389/frsen.2022.1010978
28. *Comparing land surface phenology of major European crops as derived from SAR and multispectral data of Sentinel-1 and -2.* Meroni, Michele, et al. Remote Sensing of Environment. 2021, 253, 112232. Available from: doi: http://dx.doi.org/10.1016/j.rse.2020.112232
29. *Estimating Crop Yields with Remote Sensing and Deep Learning.* de Freitas Cunha, Renato Luiz and Silva, Bruno. In 2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference. 2020, 273-278. Available from: doi: http://dx.doi.org/10.48550/arXiv.2007.10882
30. UAB Zemdirbiu Konsultacijos. Smart Agrometer. [Online] [Cited: 22 February 2026.] Available from: https://www.agronom.lt/