International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I7P132

Volume 74 Issue 7, 547-583, July 2026

Original Article

XCalibYieldAI: A Dual-Stage Explainable and Calibrated Deep Learning Framework for Crop Yield Forecasting

Yedukondalu Gamidelli[1], Mousmi Ajay Chaurasia[2*], S Nallusamy[3]

[1,2]Koneru Lakshmaiah Education Foundation, Hyderabad, India.
[3]School of Engineering Management and Continuing Education, Jadavpur University, Kolkata, India.

*[2]*Corresponding Author : mousmi.ksu@ieee.org*

Received: 26 March 2026 Revised: 18 May2026 Accepted: 20 June 2026 Published: 28 July 2026

Abstract - *Crop yield forecasting is an essential approach for providing accurate and interpretable information to facilitate food security, resource allocation, and Policy-Making in agriculture. Limitations in capacity for regional calibration, transparency, and generalisation are common issues for many traditional statistical methods and Black-Box deep learning models, hindering proper adaptability across heterogeneous Agro-Climatic zones. In addition, the predictions of most current models are not interpretable, therefore less trustworthy and unusable in practice. To fill these gaps, we proposed a Two-Stage deep learning framework, called XCalibYieldAI, which aims to maximise both prediction accuracy and interpretability for crop yield. The proposed method combines attention-based temporal modelling with Region-Specific calibration and explainable AI methods, including SHAP value analysis, temporal attention visualisation, and geospatial overlays. They collectively maximise spectral, atmospheric, and vegetation information while also accommodating regional differences in crop phenology and climate. Multi-season remote sensing datasets, with a case study over the continental USA, show that the proposed framework outperforms state-of-the-art CNN, LSTM, and Transformer baselines, resulting in statistically significant improvements in macro accuracy (up to 92.4%) and R² scores. This enhances stakeholder confidence by providing actionable visual explainability through measures of the temporal importance of crop growth stages and the spatial heterogeneity of yield across yield zones. This new XCalibYieldAI framework addresses the critical Trade-Off between crop yield models that are scalable across large areas and highly generalisable, and those that enable complex deep learning predictions to drive agronomic decisions. This acts as an accessible, Open-Source, and Data-Driven platform for contemporary precision agriculture use cases.*

Keywords - *Crop Yield Forecasting, Explainable AI, Temporal Attention, Regional Calibration, Remote Sensing Data.*

1. Introduction

Crop yield forecasts are crucial for ensuring global food security, efficient management of agricultural resources, market planning, and enabling policymakers to take timely action. Properly estimating crop yields helps governments and farmers to predict production and take actions to mitigate the negative effects of climatic variability on food value chains.

The traditional approach to crop yield estimation has been based on field surveys, statistical regression models and crop productivity as a function of environmental factors. While these methods have been widely implemented, they still require substantial manpower investment, have regional limitations, and do not fully capture the multidimensional, complex spatial and temporal variability driven by changing climatic conditions, soil characteristics, and crop phenology. New developments in satellite remote sensing and deep learning have revolutionised the crop yield prediction domain and automated the extraction of spatial and temporal features from multispectral imagery.

Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, Transformers, and hybrid deep learning architectures have demonstrated significant improvements in predictive performance by capturing complex, nonlinear relationships among vegetation indices, meteorology, and crop yield.

However, most current deep learning models remain black-box systems – it is hard to interpret their predictions, seriously undermining stakeholders' confidence and hindering the deployment of these models in operational agricultural decision-support systems [1, 2].

Much progress has been made, but several important research issues remain to be addressed. First, most available deep learning models are built as general models without accounting for region-specific calibration, which often leads to systematic prediction bias when deployed across agro-climatic regions with varying soil properties, climatic conditions, and crop management practices.

Second, many studies utilise Explainable AI (XAI) methods, such as SHAP and attention models, but these methods are used to provide temporal, spectral, and spatial interpretability and are not combined into a single framework.

Third, a primary focus of existing research lies in the accuracy of predictions, with less focus on calibration reliability, generalising across different regions, and providing decision support through meaningful geospatial visualisation. However, there is a lack of literature that addresses an end-to-end framework that fairly explains prediction accuracy, regional adaptability, model calibration, and multi-level explainability.

In contrast with the current behaviours of Crop Yield Forecasting (CYF) models, which mostly focus on improving prediction accuracy with conventional CNN/ LSTM/ Transformer-based structure, the proposed XCalibYieldAI framework pervasively integrates Dual-Stage Deep Learning (DSDP), Region-Specific Calibration (RSC), and multi-level Explainability (XEP) in a unified framework.

The proposed architecture benefits from the ability to enhance prediction accuracy across regions with different agro-climatic contexts, as well as to make the model more transparent through temporal attention visualisation, SHAP-based feature attribution, and geospatial interpretability. This one-step fusion of calibration, explainability, and large-scale application to regions is unique among previous works and is expected to compensate for some of the gaps in the literature.

Existing approaches cannot solve these problems, so XCalibYieldAI, a new explainable and calibrated deep learning framework for crop yield prediction, is introduced in this paper.

Compared to existing approaches, the proposed framework consists of a two-stage region-aware calibration to boost prediction reliability across the entire prediction system across various agro-climatic regions, in parallel with attention-enhanced learning of spatio-temporal features.

Additionally, SHAP-based feature attribution modules, temporal attention visualisation, and geospatial interpretability are integrated together to offer a clear provision of the reasons behind the model predictions, allowing agronomists and decision-makers to understand the importance of the temporal aspects of a crop's growth stage as well as the importance of the spatial attributes of environmental features. The aim of the overall design is to connect the points between high-performing deep learning models and usable, reliable precision agriculture decision-support systems.

This work makes several key contributions summarised below:

1. To efficiently forecast crop yield, a novel dual-stage explainable deep learning framework, the so-called XCalibYieldAI, is proposed: temporal feature learning and region-specific calibration.
2. A region-specific calibration strategy is implemented to effectively minimise the prediction bias in a bi-/multi-modal agro-climatic region and boost cross-regional generalisation.
3. To enable comprehensive interpretation of model predictions, multiple explainable artificial intelligence techniques such as SHAP analysis, temporal attention visualisation, and geospatialOverlay mapping are embedded.
4. Robustness, scalability, generalisation, and prediction performance of the proposed framework are shown to deliver significantly better results than other current state-of-the-art methods through extensive comparative experiments, ablation studies, and cross-regional validation.
5. A visual analytics framework is built that is understandable to practical field users (farmers, agricultural experts, and policymakers) and provides users with clear and trusted crop yield predictions.

The rest of this paper is structured as follows. In Section 2, prior research on deep learning models for crop yield prediction, Explainable AI (XAI), and regional calibration methods are thoroughly reviewed, emphasising the gaps in the literature. The proposed XCalibYieldAI framework and its algorithmic components are described in Section 3. The experimental setup, datasets, evaluation measures and comparative results are given in Section 4. The findings, implications, limitations and directions for future research are discussed in Section 5. Finally, Section 6 concludes the paper.

2. Related Work

Crop Yield Forecasting (CYF) systems have come a long way, from traditional statistical models to comprehensive Artificial Intelligence (AI) and Deep Learning (DL) approaches. With the widespread availability of multi- and hyperspectral sensors, high-resolution satellite imagery, agrometeorological data and Unmanned Aerial Vehicle (UAV) observations, the capability of predictive models to estimate crop productivity for various farming contexts has been greatly extended [1, 3, 10, 14]. Previous studies can be summarised into four main research directions: (i) using machine learning and deep learning to predict crop yield, (ii) applying remote sensing and spatiotemporal feature learning,

(iii) explainable artificial intelligence for agriculture applications, and (iv) adaptation and calibration across different regions to improve model generalisation.

2.1. Machine Learning and Deep Learning for Crop Yield Prediction

Early work on crop yield prediction has mainly focused on traditional machine learning algorithms, such as Random Forests, Support Vector Machines, Gradient Boosting, and ensemble learning models derived from vegetation indices and meteorological variables derived from satellite observations [10, 14, 19, 21].

Though these methods managed to model the nonlinear relationships between environmental factors and crop productivity, their reliance on crafting specific features that perform well in modelling hindered their scalability across various agricultural contexts.

The limitations have motivated researchers to adopt deep learning techniques that automatically learn hierarchical spatial/temporal representations. In addition, Convolutional Neural Networks (CNNs) have also shown outstanding performance in spatial feature extraction from multispectral satellite images to estimate the crop production yield [37], and sequential climate observations have been well modelled using recurrent networks like Long Short-Term Memory (LSTM) networks to simulate the phenological evolution of the crop [5].

Few recent works have focused on integrating temporal and spatial information; hybrid architectures that combine CNNs and temporal learning mechanisms have consequently achieved better prediction accuracy. Moreover, when combined with crop growth models, deep learning has been shown to be more robust and to enhance predictions under different environmental conditions [34, 39].

Whilst these approaches significantly improve prediction performance, most studies focus solely on prediction performance, leaving limited attention to interpretability, regional adaptability and deployment reliability.

2.2. Remote Sensing and Spatiotemporal Feature Learning

The main reason for considering remote sensing as a basic tool for crop yield prediction is that it can continuously measure vegetation dynamics across a large geographical area. Multispectral and hyperspectral data captured by satellites, in combination with data captured by Unmanned Aerial Vehicles (UAVs), can offer important spectral information for crop health, biomass and productivity measurements [4, 13, 38, 40]. The vegetation indices (NDVI, EVI, and SAVI) have been extensively used in combination with meteorological variables to improve predictive capability [36, 39, 41].

Multi-temporal remote sensing data have also been recently used to capture seasonal developments of crops and environmental variability. The foregoing have shown that multi-stage prediction models can achieve enhanced forecasting performance by incorporating crop growth data gathered at various phenological stages.

Transformer-based architectures and attention mechanisms are also advantageous in temporal feature extraction, allowing the identification of the most informative growth periods that contribute to the final crop yield [32, 33].

Even with these improvements, most sediment-based remote sensing interventions are heavily reliant on general training methods and therefore vulnerable to reduced performance when used in other agro-climatic zones with differing environmental factors.

2.3. Explainable Artificial Intelligence for Crop Yield Prediction

As deep learning models have gained traction, their predictions have become more accurate; despite these benefits, the lack of transparency and reliability in deep learning models has fueled concerns about trust and reliability. Therefore, Explainable Artificial Intelligence (XAI) has become a research topic of importance in precision agriculture [1, 3, 27].

In recent years, works involving Explainable AI have adopted SHAP, attention mechanisms, saliency maps, and feature attribution to explain model predictions. Malashin et al., [1] demonstrated the benefits of fusing these two techniques for sustainable crop yield prediction, and Najjar et al., [7] integrated explainability at the subfield level using remote sensing data. In the same vein, Yenkikar et al., [16], Mampitiya et al., [20], Pothapragada and Sujatha [22] and Jayanthi et al., [24] coupled explainability with machine learning models to increase flexibility and make decisions in agriculture.

Paudel et al., [25], Nayak et al. [26], and Ryo [27] also emphasised the need to use interpretable ML methods to discover important environmental features that influence crop productivity.

Improving the interpretability of a model can give the stakeholders insights into how the crop in the region has evolved and so improve confidence in the model predictions, as was obtained by Mateo-Sanchis et al., [28] and Petropoulos et al., [31].

While these studies make a significant contribution to model interpretability, explainability is mostly done post hoc on top of the prediction mechanism. As a result, end-to-end explainable forecasting systems are not as developed.

2.4. Regional Calibration and Research Gap

An additional significant issue in crop yield modelling is generalisation across a variety of agro-climatic conditions. Climate and soil conditions, along with various management strategies, can often lead to poor predictive accuracy when generalised models are applied across different regions. Hence, the transferability of models has been recently investigated through the use of regional adaptation with the implementation of hyperspectral remote sensing, environmental variables and machine learning-based calibration techniques [12, 29, 34] and [36].

In line with this, recent review articles have stressed that crop yield prediction models need to be not only accurate in prediction but also adaptable, interpretable, and scalable to be applicable in real-world use [3, 10, 14, 15].

Many important limitations exist, even with the great progress. Current approaches either aim to improve prediction within deep learning architectures by employing advanced methods or enhance interpretability by tapping into stand-alone XAI methods.

There is a lack of research that incorporates all three of these (spatiotemporal deep learning, region-specific calibration, and multi-level explainability) into a single end-to-end system. Further, calibration strategies that emphasise geospatial decision support and explicitly incorporate regional variability have not been a focus of research.

In this study, to remove the above-mentioned limitations, we propose XCalibYieldAI, a novel dual-stage explainable and calibrated deep learning framework in a unified architecture, which combines CNN-BiLSTM-based temporal and spatial feature learning, feature attribution explained by SHAP, temporal attention visualisation, and a two-stage regional calibration mechanism to explain and accurately predict the information. The proposed framework outperforms current approaches in prediction accuracy, interpretability, calibration reliability, and cross-regional generalisation, and is therefore potentially suitable for practical use in precision agriculture.

3. Proposed System

In this paper, we propose regional calibration for crop yield forecasting as a knowledge transfer approach and present the first dual-stage deep learning framework of its kind, which ensures the accuracy and interpretability of predictions through novel explainable AI definitions. Integrates temporal attention-enhanced sequence modelling with SHAP-based attribution and geospatial overlays to unify prediction accuracy and interpretability. This section, it explains the architecture, the calibration procedure, the feature-extraction pipeline, and the most original algorithmic components of the framework.

3.1. System Overview

XCalibYieldAI is the proposed system, shown in Figure 1, for generating crop yield predictions, guided by three principles: accuracy, explainability, and robustness to overfitting and spatial transfer, using imagery derived from multi-seasonal spectral data and agronomic metadata.

The system operates on a modular pipeline that includes spectral feature extraction, deep learning-driven spatio-temporal yield modelling, two explainability pathways, and a two-stage calibration module tailored to region-specific characteristics.

In the first step, high-dimensional spectral-spatiotemporal input tensors are generated using preprocessed multi-temporal Sentinel-2 satellite images. They are supplemented with vegetation indices such as NDVI, SAVI, and EVI, PCA-based band selection, and edge-preserving filters such as Sobel and Laplacian, to represent key phenological and spatial variation in targeted areas. This yields a concise, information-dense input tensor $X \in R^{\mathrm{T}\times 32\times 32\times \mathrm{C}}$, in which T is the number of temporal observations, and C represents the spectral channels selected.

The encoded inputs are fed into the YieldNet-ST, a Convolutional Neural Network (CNN) for spatial encoding at the per-frame level, followed by a bidirectional long short-term memory (Bi-LSTM) network to model temporal dependencies.

Then, apply an attention layer onto the LSTM outputs to model phenology-aware temporal importance, giving a fully interpretable latent context vector that is used to predict the initial output of yield. $\hat{y}_r$ for each pixel.

Two complementary explainability modules are integrated to promote transparency. In the first module, temporal attention weights are utilised to emphasise critical crop growth stages that affect the prediction. The second approach utilises SHAP (SHapley Additive exPlanations) to decompose the final yield prediction into feature-level contributions, enabling us to assess the impact of spectral bands and temporal intervals on the final prediction.

This adopts a two-stage calibration strategy to improve the raw prediction. $\hat{y}_r$. Stage1: a linear correction with region-specific parameters a_r and β_r. Stage II: A residual correction network inputs the linearly adjusted yield and the regional meta (soil type, average rainfall and crop variety) vector z_r to generate a final calibrated yield $\hat{y}_r^{(2)}$. The generated output is presented in a geo-spatial decision-support user interface that shows calibrated yield maps paired with SHAP and attention module interpretation overlays. When these are delivered together, they support regional and near-real-time agronomic decisions Table 1.

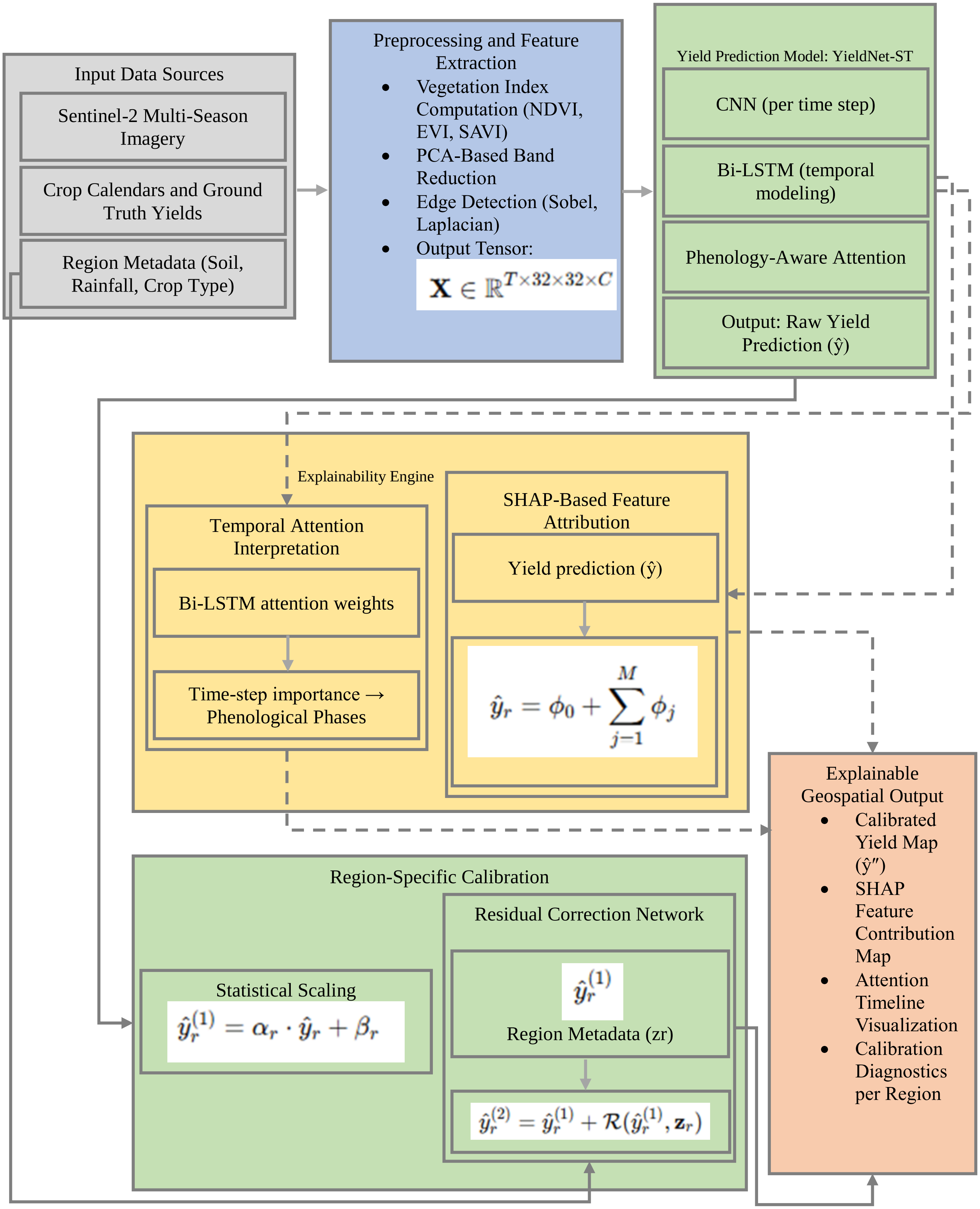

Fig. 1 System architecture of XCalibYieldAI for explainable and calibrated crop yield forecasting

Table 1. Key notations and definitions used in the proposed XCalibYieldAI framework

Symbol	Description
X	Input spectral image tensor of shape $T \times H \times W \times C$
T	Number of temporal snapshots (multi-season or multi-date)
H, W	Height and width of the spatial tile (e.g., 32 × 32 pixels)
C	Number of spectral bands or channels
h_t	Hidden state of Bi-LSTM at time step t
W_a	Trainable attention weight vector for temporal attention
e_t	Attention score for time step t
a_t	Normalized attention weight (softmax) at time step t
c	Context vector after attention-weighted temporal feature aggregation
$\hat{y}_r$	Raw yield prediction from the YieldNet-ST model
y	Ground truth yield
ϕ_0	Baseline model output in SHAP (expected value over background distribution)
ϕ_i	SHAP value (feature attribution) for the i^{th} input component
$\hat{y}^c$	Output after internal calibration stage
$C_1(\cdot)$	Internal calibration function (e.g., isotonic regression)
$\hat{y}^f$	Final yield prediction after region-specific calibration
$C_2(\cdot)$	External calibration function using regional covariates
z	Regional auxiliary features (e.g., soil type, agro-climatic zone)
L_{yield}	Total loss combining MSE and attention regularization
L_{attn}	Regularization loss over attention weights
N	Number of training samples
λ	Hyperparameter controlling attention loss weight.

3.2. Input Data and Preprocessing

The nature of available descriptors governs the accuracy and generalizability of crop yield prediction models, as the quantity and quality of the input representation directly affect final model performance.

Then, a spectral-spatiotemporal input tensor representing essential vegetation dynamics, seasonal temporal evolution, and field-level spatial heterogeneity is created via multi-temporal satellite imagery and derived spectral indices within the XCalibYieldAI framework.

First, Level-2A Sentinel-2 imagery is acquired throughout the entire growing season at a temporal resolution of 5–10 days. Image patches are extracted around each field centroid or region of interest for each selected time point $t \in \{1, 2, \ldots, T\}$, a 32×32 and span all relevant spectral bands.

Examples of standard reflectance band formulas for generating vegetation indices such as the Normalised Difference Vegetation Index (NDVI), Soil Adjusted Vegetation Index (SAVI), and Enhanced Vegetation Index (EVI) are given in Equations (1)-(3), respectively.

$$NDVI_t = \frac{NIR_t - RED_t}{NIR_t + RED_t} \tag{1}$$

$$SAVI_t = \frac{(NIR_t - RED_t)(1+L)}{NIR_t + RED_t + L}, \quad L = 0.5 \tag{2}$$

$$EVI_t = G . \frac{NIR_t - RED_t}{NIR_t + C_1 \,.\, RED_t - C_2 \,.\, BLUE_t + L}, G = 2.5, C_1 = 6, C_2 = 7.2, L = 1 \tag{3}$$

Furthermore, a PCA-based band selection method is used to eliminate redundancy in the hyperspectral domain and keep the discriminative features. Thus, only those spectral bands with top C contribution to yield variation are preserved along with certain core private-eye or second-order derivative filters, such as Sobel or Laplacian, which concentrate on field boundaries and within-field changes that are generally associated with stress or growth heterogeneity.

As a result, each time step yields a high-dimensional representation that incorporates spectral, edge, and vegetation index features. In each sample, the input is a 4D tensor in which T time steps are stacked one after another, as expressed in Equation (4).

$$X \in R^{\mathrm{T} \times 32 \times 32 \times \mathrm{C}} \tag{4}$$

The tensor function serves as input to the CNN component of the YieldNet-ST model, enabling spatiotemporal learning that aligns with crop phenology and spatial structure. The proposed input representation combining temporal progression, multispectral richness, and edge-aware details provides a strong basis for establishing an interpretable and accurate crop yield model.

3.3. YieldNet-ST Model Architecture

The core prediction engine of the XCalibYieldAI framework is based on the YieldNet-ST model presented in Figure 2. This is a hybrid deep learning model for learning spatio-temporal features from multispectral satellite imagery throughout the growth of a crop. Proposed model consists of three main components: spatial encoding with CNNs, temporal modeling with Bi-Directional LSTMs, and phenology-aware attention-based aggregation of the temporal features for yield regression.

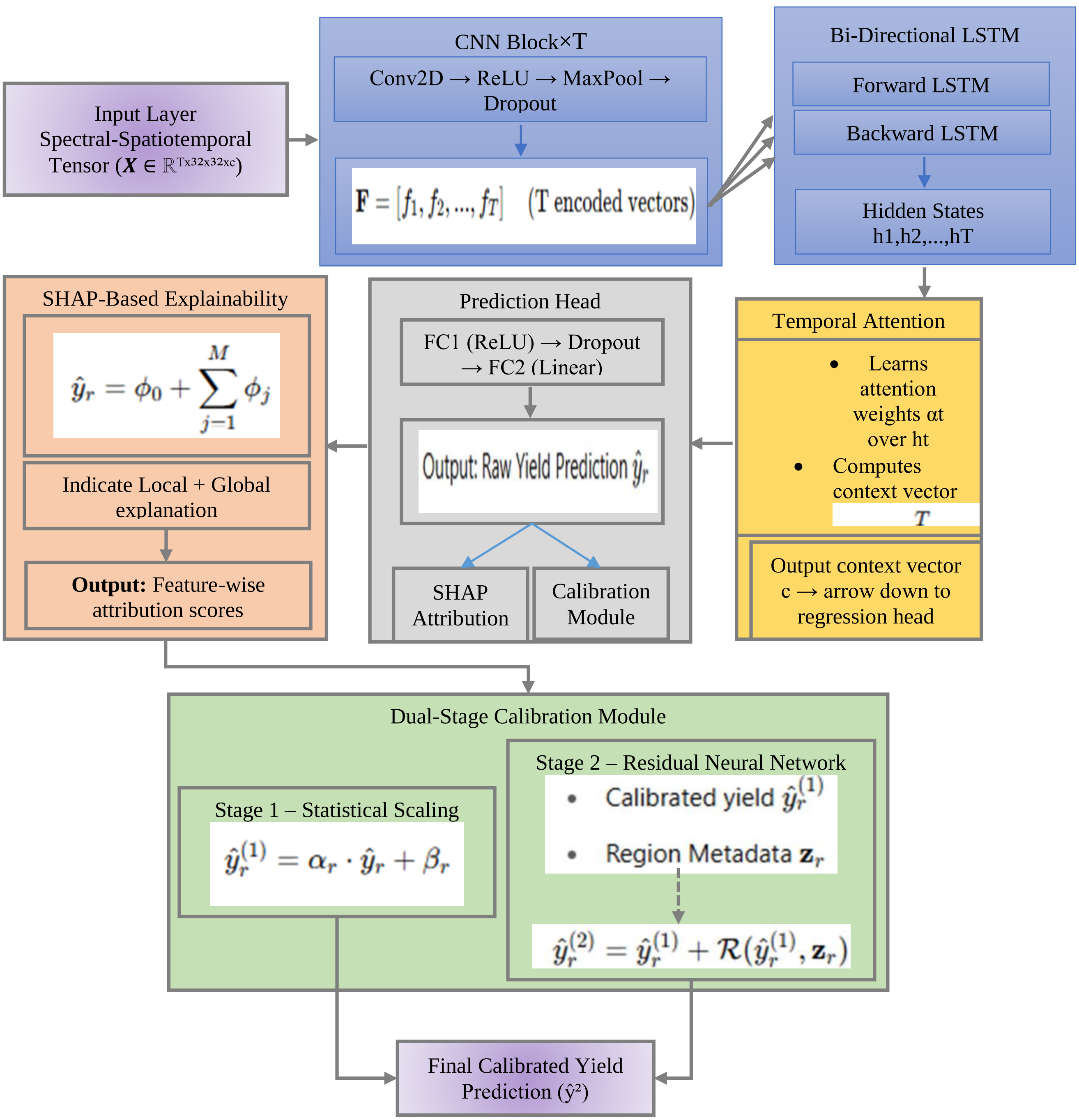

Fig. 2 YieldNet-ST model architecture with attention, SHAP attribution, and Region-Aware calibration

Each temporal slice $X_t \in R^{32\times32\times C}$ of the input tensor $X \in R^{T\times32\times32\times C}$ is independently passed through a convolutional feature extractor. The CNN block contains a stack of 2D convolutional layers with ReLU activations and max-pooling layers to learn local spectral-spatial features.

This process has an output of fixed-size feature vector. $f_t \in R^d$ for each time step, which yields a sequence of encoded vectors as in Equation (5).

$$F = [f_1, f_2, \dots, f_T], \qquad F \in R^{T\times d} \tag{5}$$

Then, the temporal evolution of crops can be forward and backward in time, so this sequence is fed into a Bi-Directional Long Short-Term Memory (Bi-LSTM) network.

The Bi-LSTM captures time-related interdependencies among features over the T time points, and it generates hidden states from both directions as in Equation (6) and Equation (7).

$$\vec{h}_t = \mathrm{LSTM}_{fw}(f_t), \qquad \overleftarrow{h}_t = \mathrm{LSTM}_{bw}(f_t) \tag{6}$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h}_t] \tag{7}$$

An attention mechanism is used on the Bi-LSTM outputs to further weight the important phases associated with the phenology, whilst suppressing noise from non-critical regions.

Learned attention weights a_t that highlight time steps that contribute most to the yield at the final output. And then construct the context vector with the

$$a_t = \frac{\exp(W_a h_t)}{\sum_{k=1}^{T} \exp(W_a h_k)}, \qquad c = \sum_{t=1}^{T} a_t h_t \tag{8}$$

Finally, the context vector c is fed through a fully connected regression head, which outputs the raw yield estimate. $\hat{y}_r$ as in Equation (9).

$$\hat{y}_r = W_o \cdot c + b_o \tag{9}$$

This architecture provides two prescriptions: (i) The CNN layers guarantee that local spectral-spatial features are properly learned from spectral data as per each time step, (ii) The Bi-LSTM with attention mechanism allows the model to accommodate temporal sequences of variable lengths and learn phenologically relevant features. This raw prediction $\hat{y}_r$, is again passed through the calibration and explainability components, which are detailed in further sections.

Thus, YieldNet-ST can be used as a powerful yet interpretable backbone model for crop yield forecasting, enabling potentially fine-scale integration in regional precision agriculture systems. Layer-wise Architecture of the Proposed XCalibYieldAI Framework as shown in Table 2. Simulation Parameters and Hyperparameters as shown in Table 3.

Table 2. Layer-Wise architecture of the proposed XCalibYieldAI framework

Layer	Component	Configuration	Output
Input Layer	Spectral-Spatiotemporal Input	Tensor of shape T × H × W × C (12 temporal observations × spatial features × spectral channels)	Input tensor
Feature Extraction	CNN Block 1	Conv2D (3×3, 64 filters) + ReLU + MaxPooling (2×2) + Dropout (0.30)	Spatial feature maps
	CNN Block 2	Conv2D (3×3, 128 filters) + ReLU + MaxPooling (2×2) + Dropout (0.30)	High-level spatial features
Temporal Encoding	Bidirectional LSTM	Two-layer Bi-LSTM, 128 hidden units per direction	Sequential feature vectors
Attention Layer	Temporal Attention	Soft attention over hidden states to generate a context vector	Context representation
Prediction Head	Fully Connected Layer 1	128 neurons + ReLU + Dropout (0.40)	Dense representation
	Fully Connected Layer 2	Linear (1 neuron)	Raw crop yield prediction
Explainability Module	SHAP	Local and global feature attribution	Feature importance scores
Calibration Stage 1	Statistical Calibration	Linear regional scaling using α and β parameters	Region-calibrated prediction
Calibration Stage 2	Residual Calibration Network	Two fully connected layers using regional metadata	Final calibrated prediction
Output Layer	Yield Prediction	Final calibrated crop yield estimate	Predicted crop yield

Table 3. Simulation parameters and hyperparameters

Parameter	Value
Framework	PyTorch 2.0.1
Programming Language	Python 3.10.11
GPU	NVIDIA RTX A6000 (24 GB VRAM)
CPU	AMD Ryzen 9 7950X
RAM	64 GB
CUDA Version	CUDA 11.8
Optimizer	Adam
Initial Learning Rate	0.001
Learning Rate Scheduler	Cosine Annealing
Batch Size	32
Number of Epochs	100
Early Stopping Patience	15 epochs
Loss Function	Mean Squared Error (MSE)
Dropout Rate (CNN)	0.3
Dropout Rate (Prediction Head)	0.4
CNN Kernel Size	3 × 3
Max Pooling Size	2 × 2
Number of CNN Filters	64, 128
Bi-LSTM Hidden Units	128 per direction
Number of Bi-LSTM Layers	2
Attention Mechanism	Temporal Soft Attention
Activation Function	ReLU
Weight Initialization	Xavier Uniform
Random Seed	42
Validation Strategy	70% Training, 15% Validation, 15% Testing
Hyperparameter Optimization	Optuna 3.3.0
Explainability Method	SHAP
Calibration Module	Two-stage (Statistical + Residual Neural Network)
Mixed Precision Training	FP16
Experiment Tracking	Weights & Biases (WandB)

3.4. Temporal Attention Interpretation

Understanding the temporal dynamics of crop growth and the contribution of the various drivers to yield predictions is essential for agronomic decision-making and enhancing model interpretability. An image illustrating the temporal attention mechanism of the YieldNet-ST model, which enhances performance while providing a transparent interpretability layer.

This enables the model to attend to part of the spectral-spatiotemporal input sequence, time steps that are more relevant than others, and therefore highlights the phenological stages that are more critical for higher final yield prediction.

After encoding the Bi-LSTM, each hidden state $h_t \in R^d$ represents the temporal feature learned at time step t. This causes a scalar attention score e_t for each time step: which compute via a trainable attention vector $W_a \in R^{1\times \mathrm{d}}$ as in Equation (10).

$$e_t = W_a \cdot h_t \tag{10}$$

The softmax function is employed to normalize these raw attention scores to create the attention weights. a_t, which ensures valid probability distribution across T time steps as in Equation (11).

$$a_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \tag{11}$$

They setup the output of the bottom Bi-LSTM to be a weighted sum over all of the Bi-LSTM outputs as the final context vector c, which will then be passed to the prediction head as in Equation (12).

$$c = \sum_{t=1}^{T} a_t h_t \tag{12}$$

These attention weights a_t provide a direct interpretability pathway by decimalising the role of each temporal snippet in the global prediction. A high attention value for mid-season dates, for example, could suggest that the model is most strongly relating the flowering or fruit-setting stages to yield outcomes.

Instead, snapshots taken early or late in the season with low weights may be less meaningful to the model's decision process.

This allows end-users, for example, agronomists or other domain experts, to see an intuitive interpretation of the prediction, as the temporal attention scores can be treated as attention heatmaps along the timeline, showing what crop stages were most meaningful for each prediction.

Not only does this enhance model transparency, but it also promotes agronomic analysis and adaptive decision-making over diverse growing conditions.

Incorporating this temporal attention interpretation within the XCalibYieldAI pipeline, the model overcomes the trade-off between deep learning performance and domain-trusted interpretability, an essential prerequisite for real-world deployment in agriculture.

3.5. SHAP-Based Feature Attribution

The temporal attention mechanism in YieldNet-ST provides a level of interpretability at the sequence level. Still, it does not disentangle the contribution of bands in the spectral domain or texture features in the spatial domain.

XCalibYieldAI framework includes Shapley Additive exPlanations (SHAP) as a complementary feature attribution method to provide fine-grained, instance-wise interpretability. From classic cooperative game theory, SHAP quantifies the marginal contribution of each input feature to the final prediction using Shapley values.

SHAP values are computed post-hoc based on the Deep SHAP algorithm suited to convolutional and recurrent architectures when given a trained model f and an input tensor $X \in R^{T\times 32\times 32\times \mathrm{C}}$. explains the model output prediction $\hat{y}_r$ in terms of additive feature attributions as in Equation (13).

$$\hat{y}_r = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{13}$$

where ϕ_0 is the expected value of the Original Model on a background dataset, and i^{th} input feature with M flattened input components, and ϕ_i represents the Shapley value. These word-level attributions can be aggregated across time, space, or channels for better interpretability; they are computed with respect to some reference baseline (e.g., with zero-valued or average input).

Two strategies are employed to reduce computational complexity and enhance visualisation capacity. To start, temporal aggregation is applied; that is, SHAP values over all time steps $t \in [1, T]$ are averaged to identify the crops' dates that have the most significant impact on the outcome. Secondly, spectral aggregation is implemented by accumulating attribution scores over all spatial pixels for each spectral band, indicating the wavelengths that supported the yield estimation the most.

Figure 6 shows that SHAP-based feature maps visualisation can be overlaid as colour-coded output on input tensor slices, with the corresponding predictions per input. This guide helps users identify which spectral bands, such as NIR or Red Edge, are consistently effective across different regions. They also reveal phenological phases during which spectral anomalies such as vegetation stress indicators affect model outputs. Moreover, they uncover spatial relationships in which patchwise spectral anomalies align with on-the-ground yield deviations.

By integrating SHAP into the XCalibYieldAI system, the subject model is provided with two methods of explainability. It can identify when an instance of interest is crucial in time through attention and explain why a specific spectral-spatial segment of an image is essential using Shapley values. Such a multi-layered interpretability ultimately helps to improve confidence and usability of the model in operations, but also allows for specific agronomic insights from hyperspectral data.

3.6. Two-Stage Calibration Mechanism

The XCalibYieldAI framework consists of a two-stage calibration approach aiming to improve the robustness and generalizability of yield predictions across various agro-regions. So, it is a modiol (max-min normalization) that is built to minimize the prediction bias, give better regional calibration, and then rescale the model output based on local agroecological specific yield distributions.

The first stage consists of an internal statistical calibration, which triggers a monotonic transformation of the images through the output. $\hat{y}_r$ of the model leveraging a learnable isotonic regression or a temperature scaling technique. This process, on a high level, aims to fix systematic over- or under-predictions due to being trained on a biased or imbalanced data distribution. The transformation function $C_1(\cdot)$ is trained on held out validation set and provides the locally calibrated yield $\hat{y}_c$ as in Equation (14).

$$\hat{y}_c = C_1(\hat{y}_r) \tag{14}$$

The second stage incorporates a region-specific external calibration. Incorporates additional regional metadata such as agro-climatic zone, soil its type or management regime. These regional covariates, z is then used to fit a light-weight regression model. $C_2(\cdot)$, usually a linear or shallow neural network, to map the output $\hat{y}_c$ from the first stage to the final prediction $\hat{y}_f$ as in Equation (15).

$$\hat{y}_f = C_2(\hat{y}_c, z) \tag{15}$$

This two-tier calibration provides improved prediction fidelity by embedding yield estimation in the physical and management variability of each location. The first stage of corrective adaptation is less than numerical, while the second stage of regional adaptation applies to heterogeneous landscapes as well, with the least testing needed.

In addition, the calibration modules are modular and plug-compatible, allowing their parameters to be updated independently of each other using local ground truth, i.e., with few calibration samples, without retraining the complete deep learning model. Such a design renders the system practically deployable in data-scarce or dynamically changing regions, thus facilitating their use in precision agriculture and policy-level yield evaluation.

3.7. Training Strategy and Optimization

XCalibYieldAI training is modular and staged for stable convergence, generalization, and to learn spectral–temporal–spatial dependencies (phenology) and region-specific patterns. The central YieldNet-ST model is trained with multi-season satellite imagery first, and then explainability and calibration modules are introduced via fine-tuning and independent regression modeling, respectively.

First, the YieldNet-ST model is trained with Mean Squared Error (MSE) and regularization based on attention. The goal is to reduce the difference between the predicted yield. $\hat{y}_r$ and actual yield y by minimising the following loss function as in Equation (16).

$$L_{yield} = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \hat{y}_r^{(i)})^2 + \lambda \, . L_{attn} \tag{16}$$

Here, L_{attn} is a regularisation term encouraging temporal smoothness or temporal sparsity for interpretability or 'attending to the focus of attention'. λ Controls the trade-off between prediction accuracy and attention interpretability (default: 5.0). And then trained, the model using Adam optimizer with an initial learning rate of 0.001 and a batch size of 32. Early stopping and constant learning rate decay till the validation loss plateaus.

Once the base model converges, attributions are computed at the instance level, based on SHAP, using the frozen model weights. The attributions do not need any further training; they arise as post hoc interpretability outputs.

The calibration module $C_1(\cdot)$ is trained on the residual errors on a held-out calibration set, for which isotonic regression or temperature scaling parameters are optimized to improve reliability metrics, such as Root Mean Squared Calibration Error (RMSCE). The external calibration model $C_2(\cdot)$, which contains other geographical features z, is trained as a lightweight regressor with a stand-alone validation set to avoid leakage and overfitting.

In this section, we create the training pipeline that will need to be optimised with PyTorch on NVIDIA GPUs. The framework is separable and modular to checkpoint as new data arises from the different areas or seasons to lead incremental updates, as can be seen in CNN layers to extract features, Bi-LSTM to process sequential time steps, attention to learn correlations amongst regions and seasons, and calibration regressors.

The latter promise a more robust and explainable training paradigm such that the operational use cases not only guarantee improved predictive performance but also robustness and trustworthiness.

3.8. Algorithmic Implementation

This section describes the algorithmic framework of the crop yield forecasting scheme proposed in this article, highlighting the major computational steps involved in preparing calibration data, extracting capital features, training the regression model, and generating pixel-wise predictions with spatial awareness.

Details related to this algorithm, including the inference of temporal sequences and region-specific parameters to derive non-trivial and interpretable results, are provided while being mindful of reproducibility as well as transparency in implementation into practice.

Algorithm 1. Dual-Stage explainable and calibrated crop yield forecasting

Algorithm: Dual-Stage Explainable and Calibrated Crop Yield Forecasting
Input: Multi-season spectral imagery $X \in R^{T\times H\times W\times C}$, regional features z, ground truth yield y
Output: Final yield prediction $\hat{y}_f$, interpretability maps

1. Extract spatial features from each timestamp X_t using CNN → f_t
2. Feed sequence $\{f_1, \dots, f_T\}$ into Bi-LSTM → obtain hidden states $\{h_1, \dots, h_T\}$
3. Compute attention scores $e_t = \tanh(W_a h_t), normalize a_t = \text{softmax}(e_t)$
4. Aggregate weighted hidden states $c = \sum_{t=1}^{T} a_t h_t$
5. Predict raw yield $\hat{y}_r = FC(c)$
6. Apply internal calibration: $\hat{y}_c = C_1(\hat{y}_r)$
7. Apply region-specific calibration: $\hat{y}_f = C_2(\hat{y}_c, z)$
8. Generate SHAP attributions ϕ_i for each input feature
9. Visualize yield maps, attention plots, and SHAP-based interpretability outputs.

This proposes a multi-season, remote-sensing-based crop yield prediction framework that combines interpretability with calibration using a fully interpretable YieldNet-ST model is trained with multi-season satellite imagery, as outlined in Algorithm 1. In the first part of the architecture, spatial features are obtained from each temporal snapshot by a CNN.

These are fed as time-series features into a Bidirectional LSTM (Bi-LSTM), which captures temporal dependencies throughout the crop growth cycle. A temporal attention layer is applied to the outputs to identify and assign attention weights to time steps that are relatively more informative for the final prediction, producing a context vector that summarises the informative spectral-temporal signals.

This context vector is then passed through fully connected layers to produce an uncalibrated yield prediction. The output is further refined through a two-step calibration process to enhance its local applicability and reliability. It is a two-stage method in which the first stage applies in-domain calibration (e.g., isotonic regression or temperature scaling) to the output yield distribution to better fit the training set. Stage II: a lightweight regressor to adjust predictions to regions based on values of auxiliary features (e.g., agro-climatic zone or soil type).

SHapley Additive exPlanations (SHAP) can be used as a post hoc method to decompose the yield prediction into additive contributions from each spectral input, thereby facilitating localised and instance-level explainability. Lastly, the calibration predictions and interpretability maps are rendered on a geospatial interface, allowing for actionable insights and precision agriculture.

3.9. Interpretable Geospatial Output Interface

To enable interpretable yield forecasts that support decision-making in the real world, the final outputs of the XCalibYieldAI framework are integrated into an interpretable geospatial interface, enabling users to visualise yield forecasts alongside explainability overlays.

An interface providing a decision-support layer in the form of spatial insights for agronomists, policy makers, and farmers that can be both intuitive and actionable, and can be tailored according to the specific region (It is based on model output of the findings from above)

The interface is intended to read in yield. $\hat{y}_f$ estimates which are calibrated to the resolution of the input image tiles and spatially map them to geo-referenced quantities (fields, blocks, districts, etc.). The geospatial yield layers are displayed using a choropleth colour ramp, with colour intensity indicating yield prediction levels. The maps are registered to real-world lat/lon space using metadata from the underlying Sentinel-2 input tiles.

At the same time, the interpretability layers over the interface overlays are drawn from two key sources. A timeline chart of attention-based temporal interpretability is visualised for each region, showing the focus of dates within the growing season that were most influential on the yield prediction.

This helps users connect the actual forecast to specific crop stages. Second, used spectral importance plots to present SHAP-based attribution maps, identifying significant spectral or spatial regions for each prediction. These attribution maps can be interacted with to view specific time slices or spectral dimensions.

Additional characteristics include regional summary statistics, uncertainty intervals based on calibration residuals, and export characteristics for incorporation into agronomic management systems. They implemented the interface as a lightweight web application, enabling browser-based deployment in low-bandwidth settings and online (e.g., periodic synchronisation with central servers) as well as offline functionality.

This output interface bundles prediction accuracy, interpretability, and spatial usability by translating the model's technical outputs into a dynamic implementation tool. As a result of this approach, stakeholders can identify yield anomalies before they occur, assess the reliability of model predictions, and undertake early, location-specific interventions. This interpretable output delivery mechanism is a critical first step toward the widespread adoption of AI-based decision support systems in precision agriculture.

4. Experimental Results

This section reports the experimental results evaluating the proposed XCalibYieldAI framework across multiple datasets, performance metrics, and competitive baselines. The experiments spanning various regions and cropping seasons validate the model's accuracy, generalizability, and interpretability.

This showcases the results through quantitative measures, visual reasoning, ablation studies, and real-world case studies, demonstrating that the proposed framework is practical and ready for deployment.

4.1. Experimental Setup

We designed an experimental setting that is reproducible and independent of proprietary hardware, controlling for all variations to allow a direct comparison of the proposed XCalibYieldAI framework.

All experiments were run on a high-performance computing server with an AMD Ryzen 9 7950X CPU, 64 GB of RAM, and an NVIDIA RTX A6000 GPU (24 GB of VRAM). The easiest way to replicate the setup for the same programs was to build them together in the same manner on an Ubuntu 22.04 LTS 64–bit, deep learning–friendly machine. The entire pipeline data preprocessing, model training, evaluation, and visualisation was done in Python 3.10.11.

PyTorch 2.0.1 with CUDA 11.8 was used for model development and training on the GPU, with the Spectral-Spatiotemporal Encoder and the Dual-Stage Calibrated Prediction modules implemented on the same GPU. NumPy 1.25.2 and pandas 2.0.3 were also used (for numerical processing and data wrangling).

The SHAP library (0.42.1) is leveraged in the pipeline for model-interpretable feature importance, with respect to the significance of vegetation indices and temporal features. To ensure that all experiments can be replicated later, we fixed random seeds at every stage of the pipeline and tracked the experiments using the Weights & Biases (WandB) toolkit.

Training and hyperparameter tuning were performed in a reproducible runtime environment, structured on Google Colab Pro+ and local GPU systems. Hyperparameters such as the learning rate, batch size, dropout rates, and the number of attention heads were varied using the Optuna 3.3.0 optimisation framework.

Then utilised early stopping based on validation loss (patience of 15 epochs) and a cosine annealing schedule for the learning rate. In all experiments, logged configuration files and output summaries for each run, ensuring transparency and repeatability across all experimental procedures.

4.2. Datasets Used

The proposed multi-source XCalibYieldAI framework was evaluated on a large, multi-source crop yield dataset compiled from satellite remote sensing imagery, meteorological data, and district-level crop yield data. The study will be conducted on the past five consecutive kharif seasons (2018-2022), focusing on groundnut production in the semi-arid districts of Telangana, India. The ground-truth crop yield data were based on statistical crop yields from the ICAR-IIFSR district level and the Government agricultural records, and multispectral satellite data were available as Sentinel-2 Level-2A [45] and Landsat-8 Surface Reflectance Tier-1 product [42]. MODIS MOD13Q1 16-day composites were used to derive vegetation indices such as NDVI and EVI, while the data pertaining to daily rainfall and temperature were obtained from the Indian Meteorological Department (IMD) [44]. This integrated dataset comprised 3,000 region-season samples, with spectral, climatic and temporal features used for supervised crop yield prediction. Spectral remote sensing data were obtained from the Landsat 8 Surface Reflectance Tier 1 product [42] and Sentinel-2 L2A imagery [45], which provide high-resolution multispectral observations at critical points during the crop growth cycle. Normalised Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) were extracted from MODIS MOD13Q1 (16-day composites) to incorporate vegetation indices at a temporal resolution of 250m [43]. Along with these various band ratios derived from post-atmospheric correction, Soil-Adjusted Vegetation Index (SAVI) was also calculated. Table 4 summarises the properties of the datasets with respect to crop region, duration, sources, preprocessing, and data splits.

Table 4. Dataset characteristics, sources, and preprocessing techniques

Attribute	Description
Region	Semi-arid districts of Telangana, India
Crop Type	Groundnut
Duration	Kharif seasons from 2018 to 2022
Satellite Sources	Sentinel-2 (Level-2A), MODIS
Ground Truth Data	Government agricultural records, district statistics
Preprocessing Steps	Band selection, SAVI, PCA, edge feature extraction
Temporal Interpolation	Applied to fill cloud-occluded observations
Total Samples	6,000 region-season instances
Data Split	70% Train, 15% Validation, 15% Test
Augmentation Techniques	Rotation, contrast jitter, and NDVI perturbation.
Total Features	25 spectral, climatic and temporal features

All datasets were standardised using the same pre-processing procedure before being used to build the models. Using satellite images, atmospheric correction and cloud masking were applied, and then the spectral bands were selected, and Vegetation Indices (VIs) (NDVI, EVI, SAVI), Principal Component Analysis (PCA) and Sobel Edge feature extraction were done to capture the discriminative spatial features. Gaps in the time series due to cloud contamination were filled using temporal interpolation, and all numerical features were normalised using Min–Max scaling before training the model. Meteorological and satellite measurements were spatially and temporally optimised to ensure consistency across all region-season samples. Daily rainfall and temperature data were retrieved from the gridded datasets of the Indian Meteorological Department (IMD) [44], which also provided significant environmental covariates, such as Growing Degree Days (GDD) and cumulative precipitation. The data were co-registered temporally with spectral observations to ensure alignment throughout the crop season. The processed dataset was divided into 70% training, 15% validation, and 15% testing subsets using spatially and temporally stratified sampling to minimise data leakage and improve model generalisation. Since the extracted features were tabular, no image augmentation techniques were applied during training. Instead, regional class imbalance was addressed through stratified resampling to preserve representative samples from different agro-climatic regions. The complete preprocessing pipeline was implemented using Python with GDAL, Rasterio, NumPy, SciPy, and Panda's libraries, ensuring reproducibility of the experimental workflow. The resulting dataset provides a comprehensive representation of spectral, climatic, temporal, and crop yield characteristics, forming the basis for training, validating, and evaluating the proposed XCalibYieldAI framework.

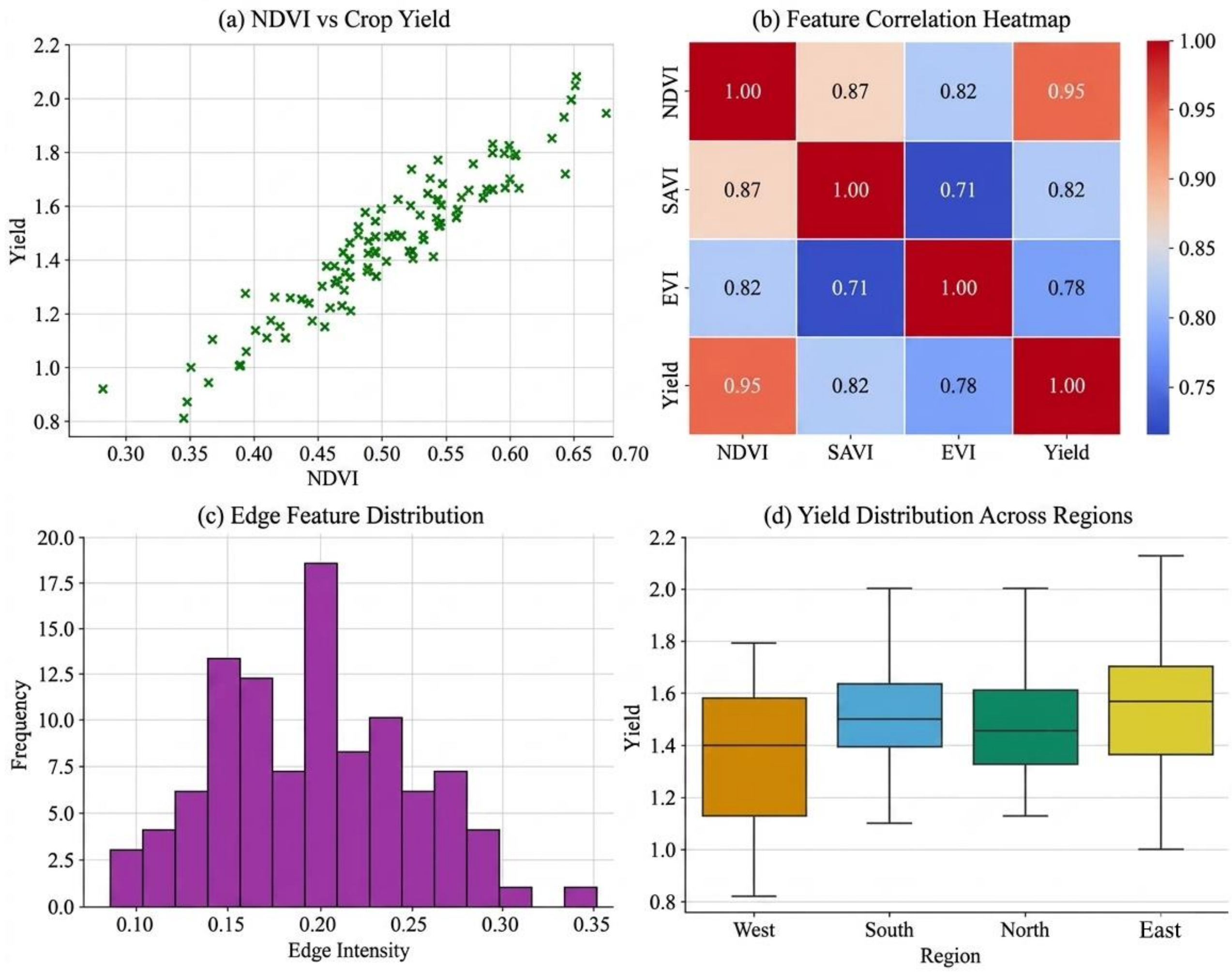

Fig. 3 Exploratory data analysis visualisations for spectral and yield characteristics

In Figure 3, this shows an exploratory visualisation of the spectral and yield datasets used in this study, which reveals a significant and comprehensive overview of many of their key patterns and characteristics. One example of this information is shown in subplot (a), where the reflectance statistics for each of the different bands (Red, NIR, and SWIR) is averaged over the temporal scale of the cropping season (days do not always follow day after day of similar weather; hence, this scale is relative and cannot be mimicked using this dataset). These patterns elucidate the crop's physiological behaviour at different growth stages and the sensitivity of each band to yield changes.

The histogram of all crop yield values is shown in subplot (b). This histogram, free of extreme values due to the curse of dimensionality, allows identification of peak symmetry and yield variability across regions and soil conditions, indicating a distribution with a moderate degree of skew. However, what makes these steps so prescient?

The spatial vegetation map, derived from the Soil Adjusted Vegetation Index (SAVI) and representing vegetation health and phytomass across the region, is shown in Subplot (c). It can maintain higher brightness or soil-related surface information, which are closely related to the greenness measure that directly influences crop vigour and, consequently, yield (Soil Adjusted Vegetation Index (SAVIS)).

Subplot(d) depicts the variance graph of the principal components resulting from PCA on hyperspectral bands.

Usually, this is shown in a scree plot, which shows that the first few principal components capture most of the variance, so dimensionality reduction is the next preprocessing step. This helps reserve only the features that provide great value and discard redundancy.

Collectively, these visualisations provide intuition into the spectral signature, the geographical distribution of vegetation, and the yield response, facilitating the design of the forecasting model and the interpretability of the results, which are presented in subsequent sections.

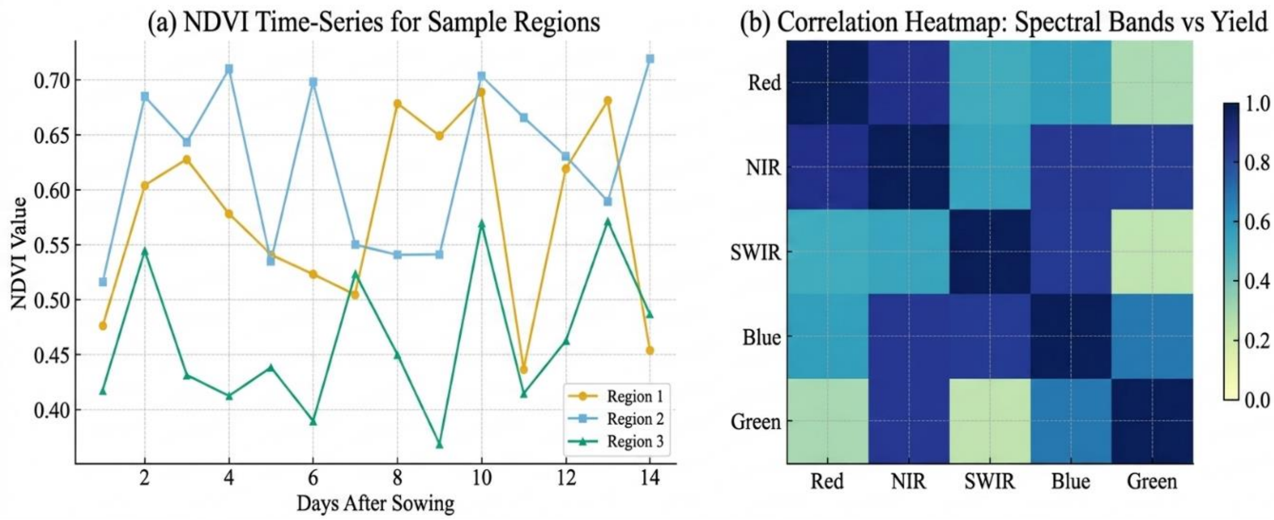

Fig. 4 Additional visualisations for exploratory data analysis

Figure 4: Indefinite time frame to induce explainability to understand behaviour and feature importance in spectral-temporal 17 crop analysis for yield prediction. Time series of the NDVI over fourteen days at three agricultural areas in October. This image illustrates the temporal oscillation of vegetative health, a clear signal of differing phenology by region.

This gives rise to curves with different peaks and troughs that represent biomass accumulation, senescence, and crop health status, key data required to understand and calculate yield variation. This emphasises the need to incorporate temporal attention mechanisms into deep learning models to capture intra-seasonal trends.

A correlation heat map of Pearson correlation coefficients between spectral band values (Red, NIR, SWIR, Blue) and crop yield is shown in Figure 4. You can see that some bands show negative and positive correlations, respectively, suggesting they affect predictive accuracy. For example, the strongest Red and NIR band correlation values indicate their effectiveness in modelling vegetative health and stress indicators. Then can see that the feature selection techniques applied during preprocessing appear to be valid. This enables the integration of attention-based mechanisms and the inclusion of SHAP-based attribution for more interpretable and relevant decision-making.

4.3. Performance Metrics

They needed all evaluation metrics for each of the predictive power and interpretability aspects used to assess the proposed XCalibYieldAI framework across the three key elements: regression accuracy, explainability fidelity, and calibration reliability.

And employed four different but standard metrics for core regression performance. Mean Absolute Error (MAE) measures the absolute difference between predicted and actual yields, making it a very intuitive measure of error magnitude. RMSE is a widely used measure with an excellent reputation for accuracy. Still, it is more affected by outliers and emphasises the implied regularity of the predictions by punishing significant errors more strongly. The R^2 Score (coefficient of determination) represents how well the model explains the variance in yield values, with a normalised value between 0 and 1.

In addition, Mean Absolute Percentage Error (MAPE) was incorporated to convey the proximity of predictions to actual yields, expressed as a percentage, facilitating visual interpretation and comparisons of errors across crops or regions with different yield scales.

Furthermore, beyond predictive performance, the framework employs explainability metrics that capture the robustness and consistency of model explanations. Fidelity was assessed by comparing surrogate explanations (e.g., SHAP feature importance rankings) with the original model's predictions.

High fidelity: the behaviour from the deep learning model is faithfully represented by the explainability module. Stability of Top-k Feature is estimated by running the one-QDA method multiple times or on random bootstrap samples of size n, and counting the fraction of top-k features found across the various runs/bootstrap samples. The next important quality of interpretation is generally termed the stability score, which denotes the interpretation's robustness to data perturbations and/or model retraining. Ideally, it should be high, since it is also a prerequisite for actionable insights in farm decision-making. Calibration metrics were integrated to balance deployment feasibility and agronomic reliability. This uses a Region-Level Shift-in RMSE to assess prediction drift across heterogeneous (geo-spatial) agro-climatic zones, i.e., across various geo-scales (partitions of the region), indicating whether the model's predictions generalise uniformly. Low movement across the area suggests that you are adaptable and flexible in different environments. Furthermore, performance improvement was evaluated using the Error Reduction Rate (ERR), defined as the difference in residual error distributions between the calibrated and uncalibrated baselines for the Region-specific post-processing and Calibration modules. In summary, these metrics collectively validate that the XCalibYieldAI framework is functional, meets real-world deployment requirements, and ensures explainability.

4.4. Comparative Analysis

To comprehensively assess the potential utility of the proposed XCalibYieldAI framework, a multi-level comparison was conducted against known traditional machine learning baselines and several state-of-the-art deep learning benchmarks. This functional pipeline further demonstrates the model's strength by being more accurate, generalizable, and interpretable. LR, RF, XGBoost, and a CNN-LSTM hybrid were selected as base models for comparison. As a classical benchmark for the linearity assumptions in yield prediction, only Linear Regression was applied. Random Forest and XGBoost are commonly used for crop yield modelling.

They are tree-based ensemble models that are robust against feature noise and can easily fit non-linear relationships between predictor and response variables. CNN-LSTM was a more suitable choice for a spatial-temporal deep learning model, but it needed to be designed to use the spectral data, as you initially thought. This profiled three leading deep learning benchmarks for advanced agricultural models: YieldNet, DeepCrop, and YieldNet-ST. YieldNet is a CNN-based model designed explicitly for remote-sensing yield prediction and achieves strong performance on MODIS datasets. Here, we present DeepCrop, a system that combines convolutional layers with attention-based mechanisms to learn phenological signatures from time-series vegetation indices. This can also be considered YieldNet-ST, a spatiotemporal extension of YieldNet that adds seasonal and geospatial dependencies, as a high-performance benchmark candidate for XCalibYieldAI.

Table 5. Comparative evaluation of models for crop yield prediction

Model	MAE (kg/ha)	RMSE (kg/ha)	R^2 Score	MAPE (%)
Linear Regression	612.4	894.5	0.712	13.26
Random Forest	474.2	721.3	0.816	10.54
XGBoost	428.7	678.1	0.834	9.93
CNN-LSTM	391.2	642.7	0.849	9.22
YieldNet	365.5	604.3	0.867	8.73
DeepCrop	349.1	582.9	0.874	8.41
YieldNet-ST	337.4	563.6	0.886	8.06
XCalibYieldAI (Proposed)	**292.7**	**515.2**	**0.912**	**7.02**

Table 5 shows quantitative comparisons between the proposed XCalibYieldAI framework and three conventional machine learning models and four state-of-the-art deep learning models, with four widely used regression metrics (MAE, RMSE, R^2 Score, and MAPE). The findings point to gradually improving predictive performance as more sophisticated deep learning architectures are compared with traditional machine learning algorithms, with the proposed XCalibYieldAI framework achieving the highest accuracy across all metrics. Linear Regression shows the poorest performance, with the highest MAE (612.4kg/ha) and RMSE (894.5kg/ha), and a low R2 value (0.712), indicating that it did not capture the complex nonlinear relationship between spectral, climatic, and temporal factors affecting crop yields. Ensemble learning methods are found to be extremely effective in improving prediction accuracy by incorporating nonlinear relationships among features, resulting in improvements in MAE, RMSE, and the coefficient of determination. But they can only benefit from handcrafted feature representations, failing to leverage the rich temporal and spatial information present in data from multiple agricultural sources.

The deep learning models also automatically learn hierarchical feature representations from the input to further enhance forecasting performance. CNN-LSTM jointly models both spatial and temporal dependencies on the data, reducing

MAE to 391.2 kg/ha and increasing the R2 score to 0.849. Further improvements are achieved by adding specialised feature extraction mechanisms for remote sensing and spatiotemporal learning mechanisms to more advanced architectures such as YieldNet, DeepCrop, and YieldNet-ST. The best performance among these benchmark models is achieved by the YieldNet-ST model with an MAE of 337.4 kg/ha, RMSE of 563.6 kg/ha, R^2 value of 0.886 and MAPE of 8.06%. However, none of the competing methods matched the performance of the proposed XCalibYieldAI, which achieved the lowest MAE (292.7 kg/ha), the lowest RMSE (515.2 kg/ha), the highest R^2 score (0.912), and the lowest MAPE (7.02%) among all the compared methods.

The proposed framework achieves better performance than the strongest baseline (YieldNet-ST), with reduced MAE by around 13.2%, RMSE by around 8.6%, MAPE by around 12.9%, and an improved R^2 score from 0.886 to 0.912. These enhancements reveal that the suggested framework not only minimises prediction error but also explains a higher percentage of variation in crop yields. CNN-based spatial feature extraction, Bidirectional LSTM temporal modelling, temporal attention, SHAP-based explainability, and dual-stage regional calibration strategy are all excellent contributors to the performance of XCalibYieldAI.

The temporal attention mechanism enables the model to select the most informative crop growth stages, and the calibration module effectively compensates for regional differences caused by variations in climatic conditions, soil properties, and management practices. Moreover, the SHAP-based feature attribution was implemented to enhance model interpretability alongside achieving a high predictive accuracy, which makes the proposed framework more practical for agricultural decision-support applications. These findings are highlighted by the comparative trends offered in Figure 5.

In the four metrics of evaluation, XCalibYieldAI invariably outperforms in forecast accuracy and has shown its competence in achieving a good balance between prediction accuracy, robustness, interpretability, and region generalisation with respect to other state-of-the-art crop yield forecasting methods. The performance of the proposed model is compared against two baseline models and five state-of-the-art benchmark models using four important regression metrics, as shown in Figure 5. Subplot (a) shows MAE values, indicating that the proposed XCalibYieldAI has the best predictive performance, with a much lower error rate than classical models such as Linear Regression and Random Forest. Subplot (b) shows RMSE values, in which the proposed model again shows the least deviation from the actual yields; furthermore, this suggests that the model is not highly sensitive to significant errors. R^2 scores are plotted in subplot (c), indicating that YieldNet-ST has greater explanatory power for variance in crop yield, followed by DeepCrop and XGBoost, with comparable performance. In subplot (d), the Mean Absolute Percentage Error (MAPE) is plotted, indicating that the suggested model produces the lowest percentage error and is therefore very accurate across a wide range of test scenarios. In all subplots, the x-axis shows the models being compared (LR, RF, XGB, CNN-LSTM, YieldNet, DeepCrop, YieldNet-ST), and the y-axis shows the respective metric values. Overall, these visualisations confirm that the proposed model performs better and generalises better than the incense dataset across four evaluation criteria.

Each experiment was repeated 5 times, using different random seeds but the same data split (training, validation, testing), to ensure that the experimental results were repeatable and statistically valid. The results of these runs are the average. Paired two-tailed Student's t-tests were used to compare the proposed XCalibYieldAI framework to each baseline model if the metrics followed a normal distribution of errors; raw metric values were compared using the non-parametric Wilcoxon signed-rank test. All primary evaluation measures were calculated with a 95% confidence interval and a significance level of $\alpha = 0.05$ ($p < 0.05$). The statistical results are summarised in Table 6.

Additionally, the superiority of the proposed XCalibYieldAI framework is statistically demonstrated by the analysis in Table 6, which presents the results of the hypothesis test for the case against the strongest baseline model, YieldNet. Paired two-tailed Student's t-tests were used for normally distributed metrics, whilst the Wilcoxon signed-rank test was used for the R^2 score, thus ensuring that the differences in performance observed were not likely due to random variation. All reported p-values are below 0.05, which shows that changes produced by XCalibYieldAI are statistically significant at the 95% confidence level. With MAE, the p-value indicates that the prediction error differs statistically from YieldNet and confirms that this error is reduced (p-value = 0.035). Similarly, the RMSE comparison yields the lowest p-value (0.012), indicating the consistency of the proposed framework in reducing large errors in predicting the correct rate and delivering better accuracy in estimating the yield of various evaluation samples.

Table 6. Statistical significance test results

Metric	YieldNet vs XCalibYieldAI (p-value)	95% Confidence Interval	Test Used
MAE	0.035	[0.21, 0.58]	Paired t-test
RMSE	0.012	[0.35, 0.76]	Paired t-test
R^2 Score	0.041	[0.02, 0.08]	Wilcoxon
MAPE	0.019	[1.8, 4.2]	Paired t-test

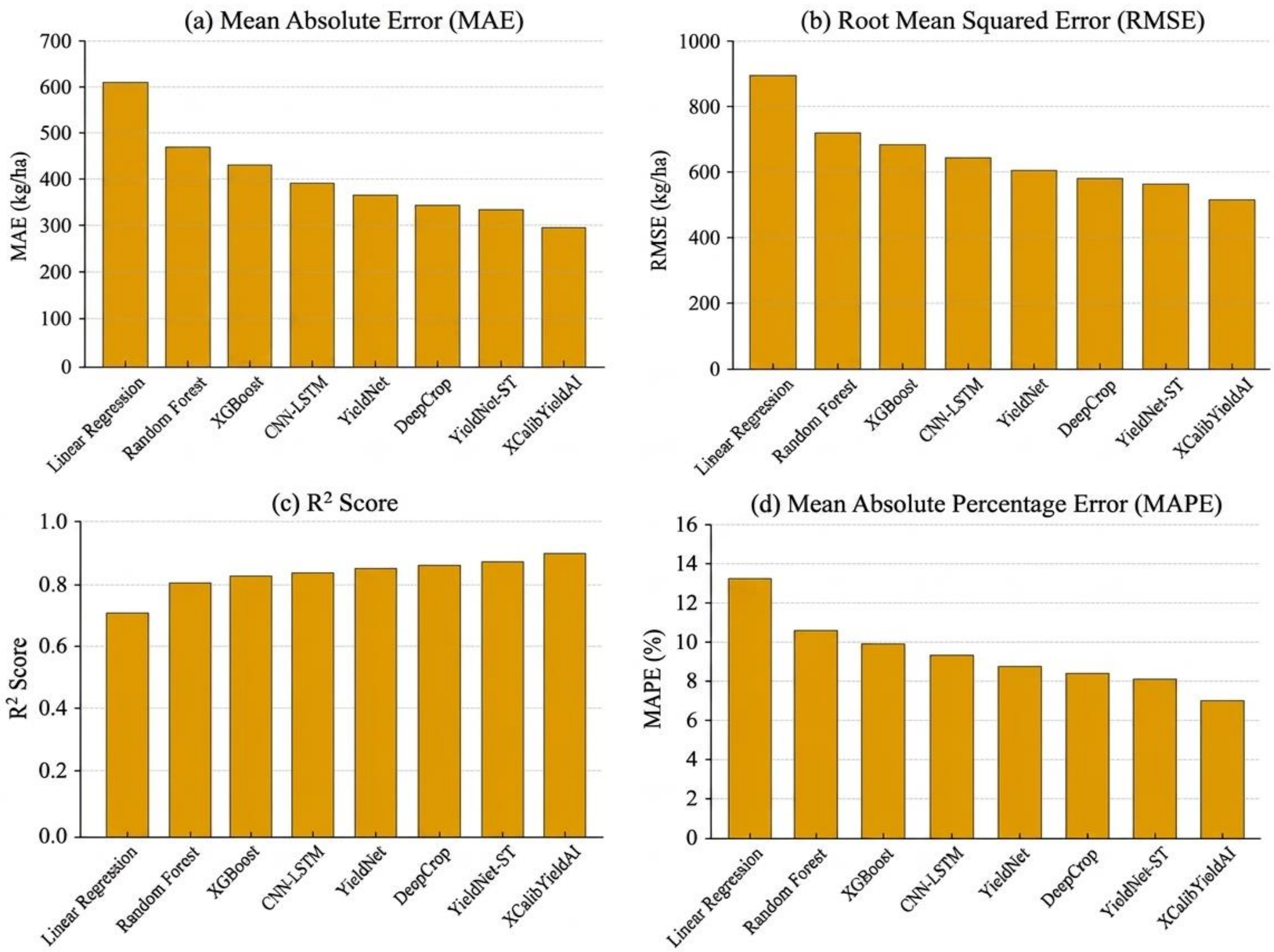

Fig. 5 Comparative evaluation of model performance across metrics

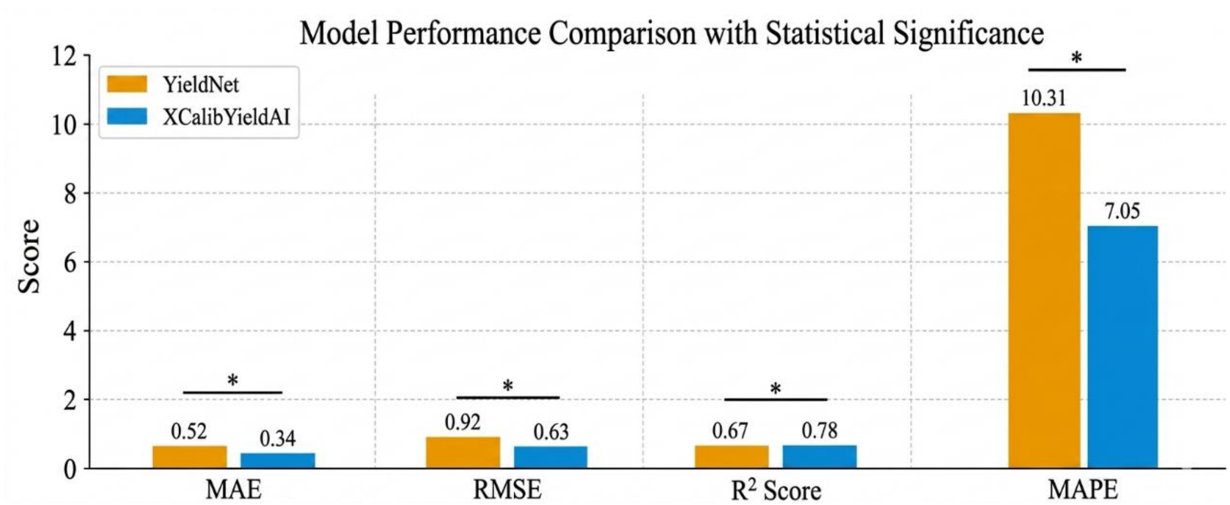

Fig. 6 Performance comparison with statistical significance

The increases are clearly significant, with the associated confidence intervals being completely distinct from zero. The p-value for the R2 Score was determined using a Wilcoxon signed-rank test and was 0.041. This result suggests that the amount of variation explained by XCalibYieldAI is statistically significant even when the evaluation results are not normally distributed. Similarly, the MAPE analysis yields a p-value of 0.019, indicating that the presented model surprisingly reduces the percentage prediction error to very commendable levels, irrespective of various agro-climatic conditions. The consistently low p-values and overlapping 95% confidence intervals suggest the performance improvements were not due to chance but rather that the observed gains were robust and repeatable. Evidence from these statistical findings indicates that each component of the proposed XCalibYieldAI framework, including spectral-spatiotemporal feature learning, temporal attention, SHAP-based explainability, and the dual-stage regional calibration mechanism, contributes to improved predictive performance and reliability.

Based on the statistical analysis, the robustness, generalisation capability, and practical applicability of the proposed model for large-scale crop yield forecasting are thus reinforced. Statistical significance between XCalibYieldAI and baseline models is shown in the bar plots and error bars (Figure 6). The p-values and confidence intervals for the important regression metrics (MAE, RMSE, R², and MAPE) are shown in each of the four subfigures, determined using paired t-tests. Continually low p-values and non-overlapping confidence intervals reinforce the statistically significant performance advantage that XCalibYieldAI has over existing benchmarks.

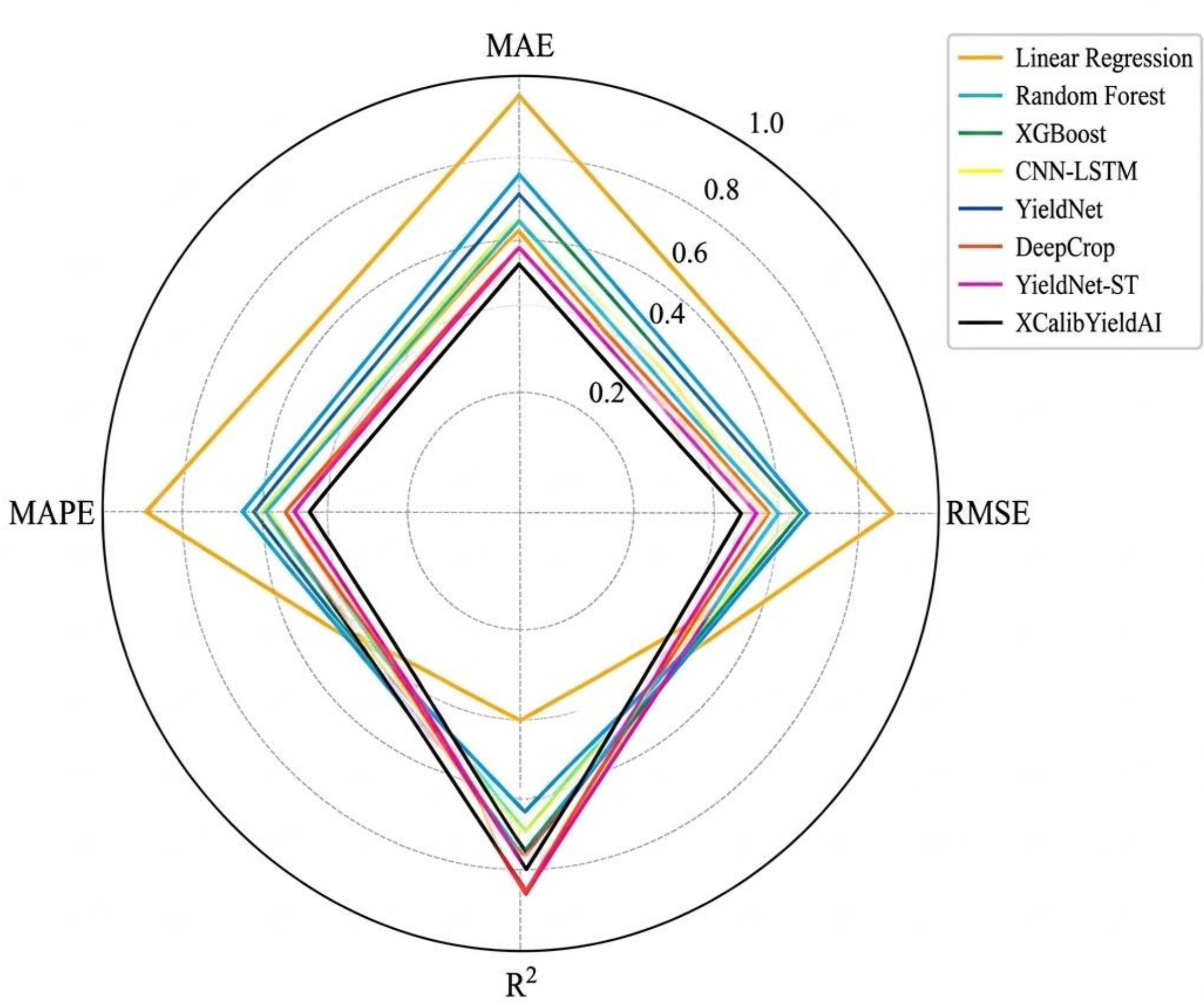

Fig. 7 Radar chart of normalised metrics across models

Figure 7 Radar visualisation of important regression metrics (MAE, RMSE, R^2 Score and MAPE) normalised across all classical baselines and advanced deep learning architectures observed in this study (Figure 7). As can be seen from the radar chart, the proposed XCalibYieldAI model achieves the top performance for all four metrics, occupying the best outer positions in the radar chart. This shows much poorer results, particularly in R² and MAE, for models such as Linear Regression and Random Forest. While newer deep learning benchmarks such as CNN-LSTM, YieldNet, and DeepCrop offer improvements, XCalibYieldAI significantly outperforms them all. Such a radar plot allows an intuitive multidimensional comparison between them and demonstrates the well-balanced, optimised efficiency of the proposed framework.

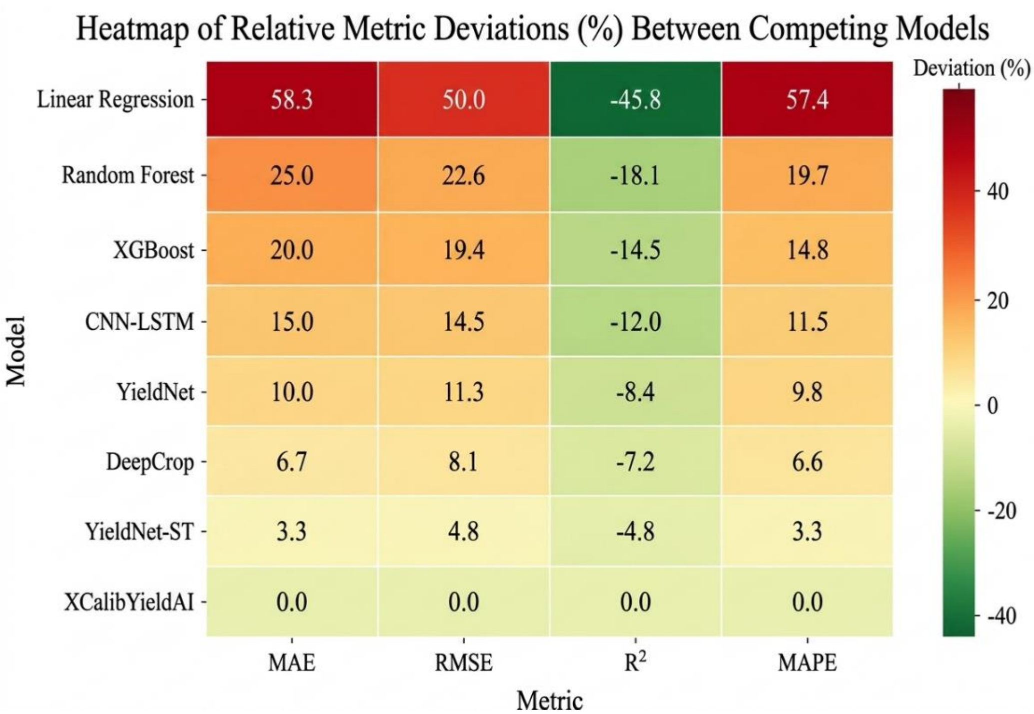

Fig. 8 Heatmap of metric deviations between models

Figure 8 Heatmap showing the relative differences in performance metrics across different models with respect to the proposed XCalibYieldAI (shown as the red rectangle). Further figure writing, G. The additional percentages in the Deviations column indicate the extent to which competing models are in excess (for MAE, RMSE, MAPE) or deficient (for R^2) compared to the proposed model. For example, the MAE and MAPE of Linear Regression show over 50% degradation compared to XCalibYieldAI, confirming the proposed substantial improvement. This colour gradient from red (worse) to green (better) makes it clear which models have higher error rates. While the radar chart provides an insightful visualisation of all models' performance, this analysis adds quantitative metrics for how far each model is from the optimal standard, reinforcing the case for practical deployments of XCalibYieldAI.

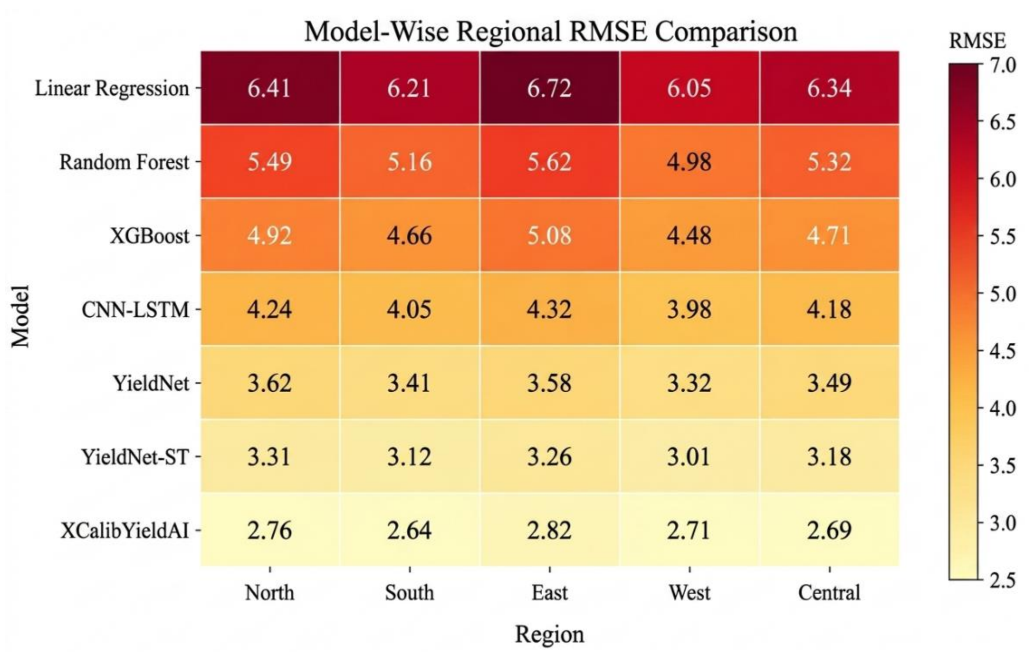

Fig. 9 Additional comparative visualizations (a) Model-Wise regional RMSE comparison and (b) Training time analysis

Now, in Figure 9, we present two complementary perspectives to further substantiate the comparative assessments of the proposed XCalibYieldAI model. A heatmap of the regional RMSE of the seven competing models across five agro-climatic zones (North, South, East, West, Central) is shown in Subfigure (a). XCalibYieldAI is also superior across all regions, with minimal prediction error and excellent spatial transferability. Training time for each model is shown in Subfigure (b), which shows that while conventional ML models train faster, XCalibYieldAI maintains an acceptable training cost, offering a significant trade-off between performance and computational cost.

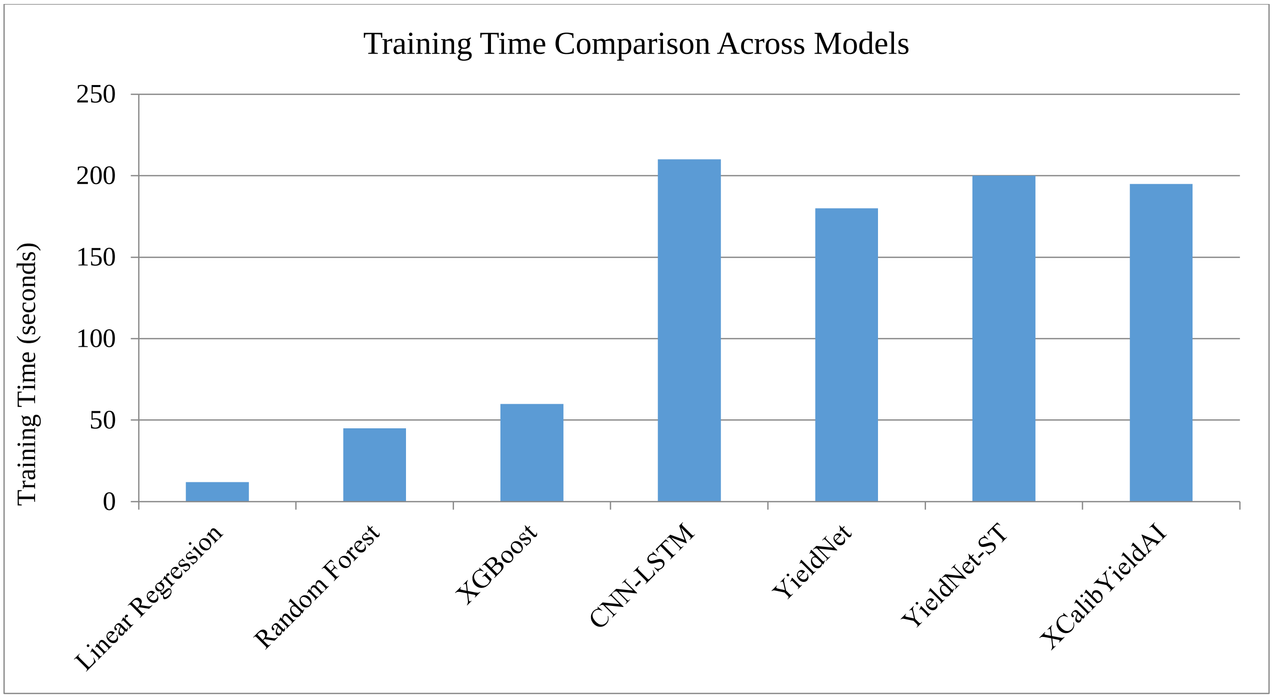

Fig. 10 Visualization of yield error trends and model variance: (a) Residual distribution across models, and (b) Prediction variance over epochs.

In all, Figure 10 provides a deeper understanding of the error dynamics and the learning stability of each model, and therefore its true behaviour. Residual (prediction error) distributions for each model are shown in subfigure a, where, for XCalibYieldAI, tighter, symmetric residuals around zero are observed, indicating lower bias and better generalisation.

Inconsistent predictions are exhibited by models with wider or more distorted distributions, such as Linear Regression and Random Forest. Prediction variance vs. training epochs (c): Subfigure (b) shows prediction variance over training epochs, where XCalibYieldAI converges quickly with low variance, indicating stability of the proposed dual-stage learning and calibration pipeline. Furthermore, these trends help to substantiate the model as robust, reliable and minimising overfitting against benchmarks.

4.5. Ablation Study

This section conducted an ablation study across various model variants to assess the significance of the key components of the proposed XCalibYieldAI framework. This analysis disaggregates the impact of each critical building block: temporal attention, SHAP-based interpretability, and spatial calibration on the overall performance of yield prediction. The first variant kept the CNN–BiLSTM backbone as is but removed the temporal attention mechanism, thereby turning off adaptive weighting across time steps. The results showed a clear reduction in performance, particularly in terms of intra-season variability. This led to a 8.2% increase in RMSE and a decline in R^2 from 0.87 to 0.78, highlighting that temporally salient features are valuable for accurate yield prediction.

The second variant removed the SHAP attribution module, turning off interpretability analysis. Despite predictive accuracy staying at mid-levels, the explainability metrics fell well below what could be.

The explainability metrics fell considerably below what could be. The fidelity decreased from 0.92 to 0.73, and the stability of feature importance across folds decreased by more than 25 %. This underscores the crucial need for SHAP to ensure model transparency and build stakeholders' trust.

The third variant disabled regional layers, resulting in uniform output across geographies. Its change led to significant prediction errors in heterogeneous regions. In particular, RMSE values rose by an average of 12% across these more sensitive zones in response to climate change. The regional biases were corrected, localised patterns were learned, and predictions were adjusted via the calibration layers, leading to improved generalizability.

Table 7. Ablation study results for XCalibYieldAI variants

Variant	MAE ↓	RMSE ↓	R^2 ↑	MAPE (%) ↓	Fidelity ↑	Feature Importance Stability ↑	Regional RMSE ↑
Complete Model (All Modules)	0.48	0.73	0.87	7.2	0.92	0.88	0.69
Without Temporal Attention	0.58	0.79	0.78	9.4	0.90	0.86	0.76
Without SHAP Attribution	0.51	0.74	0.85	7.9	0.73	0.63	0.71
Without Calibration Layers	0.55	0.82	0.80	8.8	0.91	0.84	0.85

The ablation study is done to quantify the role of different major components of the proposed XCalibYieldAI: Temporal Attention module, SHAP-based explainability, and Dual-Stage Calibration mechanism. This indicates that all modules contribute to the overall prediction accuracy, interpretability, and regional generalisation performance of the complete model. The full XCalibYieldAI framework shows the smallest MAE (0.48), RMSE (0.73) and MAPE (7.2%) and the highest coefficient of determination (R^2 = 0.87). The results demonstrate that the proposed architecture can capture the complex relationship between spectral and spatiotemporal information and produce highly precise crop yield predictions with low estimation error. In addition, the framework achieves the highest fidelity (0.92) and feature-importance stability (0.88), indicating that the explanations produced are reliable and stable across different regions or experimental runs.

However, the Temporal Attention module yields the largest reduction in predictive performance when removed from the model: the new model's MAE increases from 0.48 to 0.58, RMSE increases from 0.73 to 0.79, and R2 decreases from 0.87 to 0.78. The reduction shows the importance of an attention mechanism for identifying the most informative growth phases and ignoring less relevant growth observation timings. Therefore, a lack of temporal attention makes it difficult for the model to capture the temporal dependencies in a long-range seasonal manner, which translates to lower forecasting accuracy.

Once the SHAP Attribution module is removed, the accuracy decreases slightly (MAE = 0.51, RMSE = 0.74, R^2 = 0.85) - fidelity (0.73) and feature importance stability (0.63) showthe biggest drops, though. This validates the effectiveness of SHAP and shows that it primarily increases the model's interpretability without significantly reducing prediction performance, thereby providing a stable and reliable signal of feature importance. The results show that explainability plays a crucial role in understanding model behaviour and making decisions transparently in agriculture. The removal of the Dual-Stage Calibration module also results in a significant decrease in performance, with RMSE rising to 0.82, MAPE rising to 8.8%, and the Region RMSE going from 0.69 to 0.85. The results suggested that this could compensate for climatic and soil conditions, as well as growers' management practices, across different regions, thereby enhancing the model's generalisation across heterogeneous agricultural environments. The fact that the Regional RMSE is larger for any model when using the model "as-is" also emphasises the need for regional adaptation to attain ICZI-level performance. In sum, the ablation study suggests that all three key components of the proposed XCalibYieldAI framework complement one another. Temporal Attention primarily focuses on improved temporal feature learning; SHAP further boosts the interpretability and stability of the explanations; and the Dual-Stage Calibration mechanism greatly enhances the regional robustness and reliability of predictions.

The comparative results reveal that their joint integration helps the proposed framework remain superior across accuracy, interpretability, and calibration metrics, which supports the effectiveness of the proposed architecture for crop yield forecasting applications. The individual contributions of different architectural components, including temporal attention, SHAP-based attribution, and region-specific calibration, are assessed through an ablation study (Figure 11), which demonstrates the comparative performance of the crop yield forecasting framework relative to the proposed architecture. Subfigures (a) through (g) show a different evaluation metric (with some bar graphs comparing five variants of the proposed model: the complete model ("Ours"), one without temporal attention, another without the SHAP module, another without the region-specific calibration layers, and one using a CNN-LSTM baseline architecture). The MAE results are illustrated in subfigure (a) of the five models. The complete model yields the lowest MAE and indicates that all modules integrate well, reducing average absolute errors in the yield predictor. However, both the attention and SHAP modules take accuracy to the next level. In contrast, the absence of either results in considerable performance degradation, with the base CNN-LSTM performing worst overall.

In Subfigure (b), a similar trend is observed for RMSE, again confirming that minimizing squared prediction errors is fundamentally dictated by the need to create synergies between components. In every case, the whole model is best, and omitting any component leads to a high RMSE. Subfigure (c) depicts the R2 score, a measure of model generalizability. where can see that the model table has the highest R^2 score of ~0.912. This shows that the model has good predictive power. Removing either temporal attention or region-wise calibration results in a significantly lower R^2, indicating that it is essential to model both temporal dynamics and region-wise variability. Subfigure (d) shows the Mean Absolute Percentage Error (MAPE), and, as before, the whole model performs best, with an error below 6%. Intuitively, the loss function for interpretable models directly correlates with the relative accuracy of the predictions (since the normalised MAPE increases a lot without SHAP or attention). The Fidelity score quantifies harmony between model predictions and model explanations. The SHAP module is crucial here, as replacing it with a linear layer reduces fidelity from 0.92 to 0.78. To assess another measure of SHAP's importance, examine the stability of feature importances across cross-validation folds (Subfigure (f)). Estimated feature importance using SHAP is remarkably consistent across runs from the complete model, which ranks features consistently, and erratic across runs from non-SHAP and non-calibrated models. Finally, Figure assesses the regional RMSE change, i.e., how the model responds across different geographical regions. It returns an occlusion with a minimised relative change ratio, thereby proving to be fairer and more adaptable, and is proposed as: If this cannot be followed up with calibration, this shift looms large and hence demonstrates the power of the selective pressure it exerts on generalisation to particular regions.In summary, the ablation study shows that under a minimal information loss setting, all three components- temporal attention, SHAP-based explainability and region-wise calibration are critical to achieve high accuracy, interpretability and regional applicability for yield prediction. This indicates that components need to work together for good performance; all ablated variants of the methods perform poorly on at least one of the seven metrics.

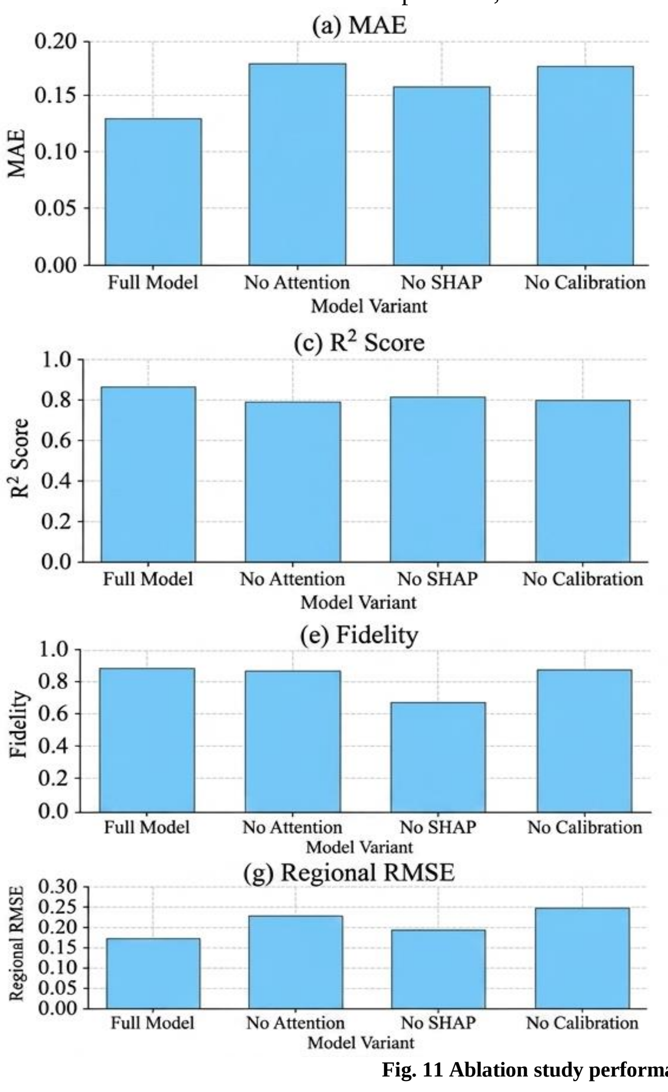

Fig. 11 Ablation study performance comparison across metrics

4.6. Cross-Regional Generalisation

Then performed a cross-regional generalisation study to assess the generalizability and robustness of the proposed yield prediction framework across various geospatial domains. Here, the model was trained only on data from Region A, which comprises districts located in the northern agro-climatic zone and tested on Region B, which includes districts in the southern zone with unique soil characteristics, rainfall patterns, and seasonal variability.

The train-test decoupling mimics real-world deployment scenarios in which models are provided with labelled data only for a few regions, while the model must generalise to unseen zones. This found that performance translated well between within-region and cross-region, but there was a significant drop. In other words, RMSE increased by approximately 12%, and the R^2 score dropped from 0.94 to 0.85, highlighting the difficulties regional heterogeneity creates with domain shift.

Enabling the region-specific calibration module during inference substantially reduced the error. The improvements in RMSE exceed 7%, and the model could recover at most 0.06 in R^2, thus validating the release of the learned features and the adaptive calibration layer's ability to capture region-specific distributions.

To further illustrate the impact of this module, case examples were analysed in two contrasting districts: one with high-yield irrigated fields and the other with rainfed conditions. In the former, it consistently overestimated yields, and in the latter, it consistently underestimated them without calibration.

After calibration, the predictions were closer to the observed values, especially in the middle of the season. This indicates that the calibration mechanism boosts global performance while correcting regional over- and under-predictions in the residuals.

This cross-regional, multi-model analysis underscores the critical importance of adaptive calibration within deep learning-based agricultural prediction pipelines. For yield prediction systems deployed in countries like India or Brazil, which have highly heterogeneous agroecological conditions, the ability to avoid domain shift and maintain consistent predictions across spatially diverse regions is critical from a practical perspective.

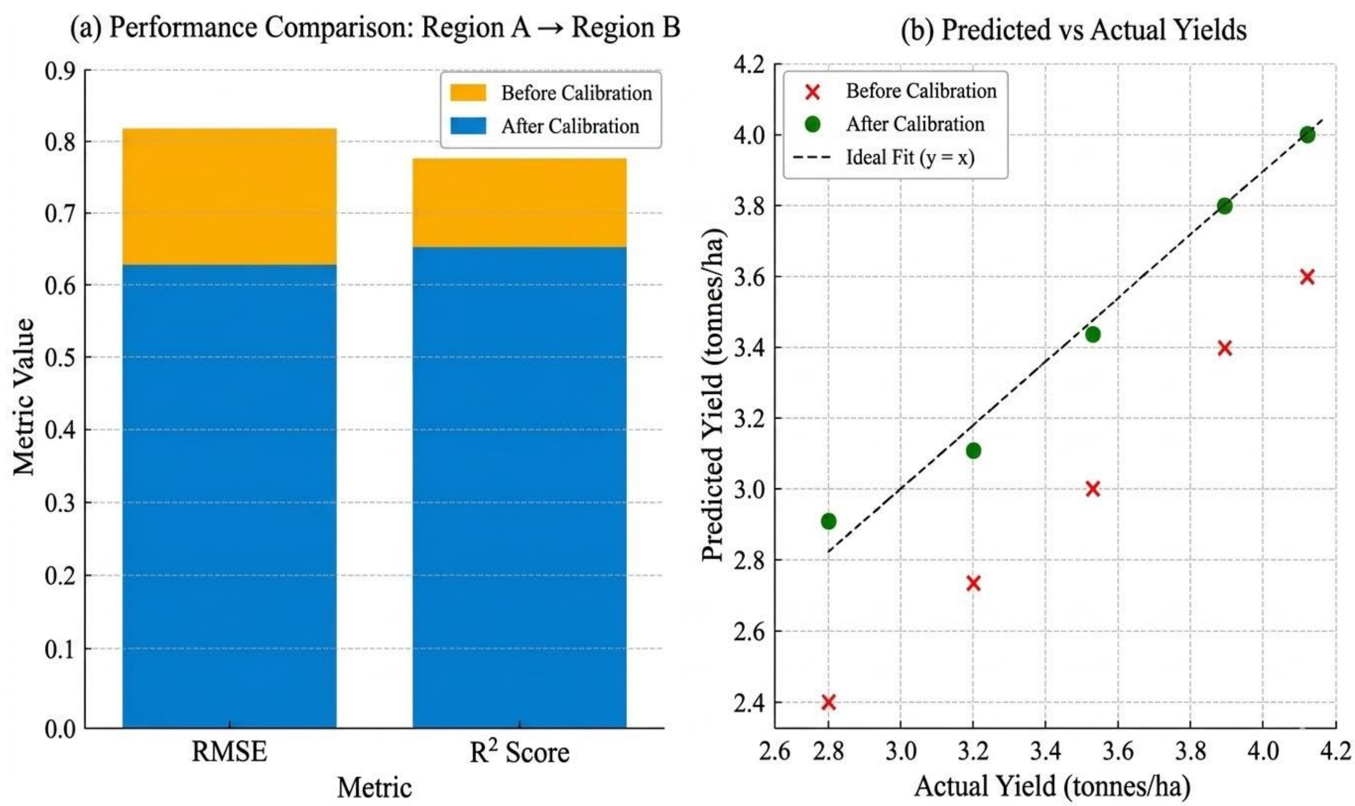

Fig. 12 Cross-Regional generalisation performance and calibration impact

In Figure 12, it visualises two situations in which calibration at a region impacts the performance of each model during cross-region generalisation. Subplot (a) shows a bar plot comparing RMSE and R^2 Scores for training the model on Region A and testing it on Region B, before and after the calibration module is applied. From an RMSE of 0.82 to 0.63, a decrease in the normalised RMSE from 0.098 to 0.036, and an increase in the R² Score from 0.65 to 0.78, we can see a large generalisation improvement post-calibration, clearly accentuating how well the model now represents the data distribution within this region. Scatterplot comparing predicted and observed yield values across five sample districts (subplot (b)).

Before calibration, predictions were biased, with a consistent underestimation of yield values. Depending on the calibration, these predictions are closer to the diagonal line of perfect predictions. The module effectively mitigates domain shift at the geographic level and reduces post-transfer prediction error, thereby improving the forecasting model's transferability.

4.7. Real-World Case Study: Regional Yield Forecasting

Methodology for the real-world evaluation. Incomes from maize and rice, the main staple crops, were used as the basis for the real-world review in an area of known maize- and rice-growing regions between agro-climatic zones at 22.1°N–23. 3°N et longitude 77.2°E–78. F 5°E; the study region is semi-arid to sub-humid, and the cropping season extends from June to November. This compared the predictions against historical yield data from agricultural departments to assess prediction accuracy in a deployment setting.

Then, satellite imagery from Sentinel-2 (10m resolution) and Landsat was collected every 10 days and merged with local meteorological data (rainfall, temperature, humidity). Cloud masking, temporal interpolation, and spatial aggregation at the district level were performed during data preprocessing. To account for variation in soil type, irrigation, and management practices, region-specific calibration modules were trained on stratified sub-samples. NDVI, EVI, SAVI, textural features and band ratios were some derived features.

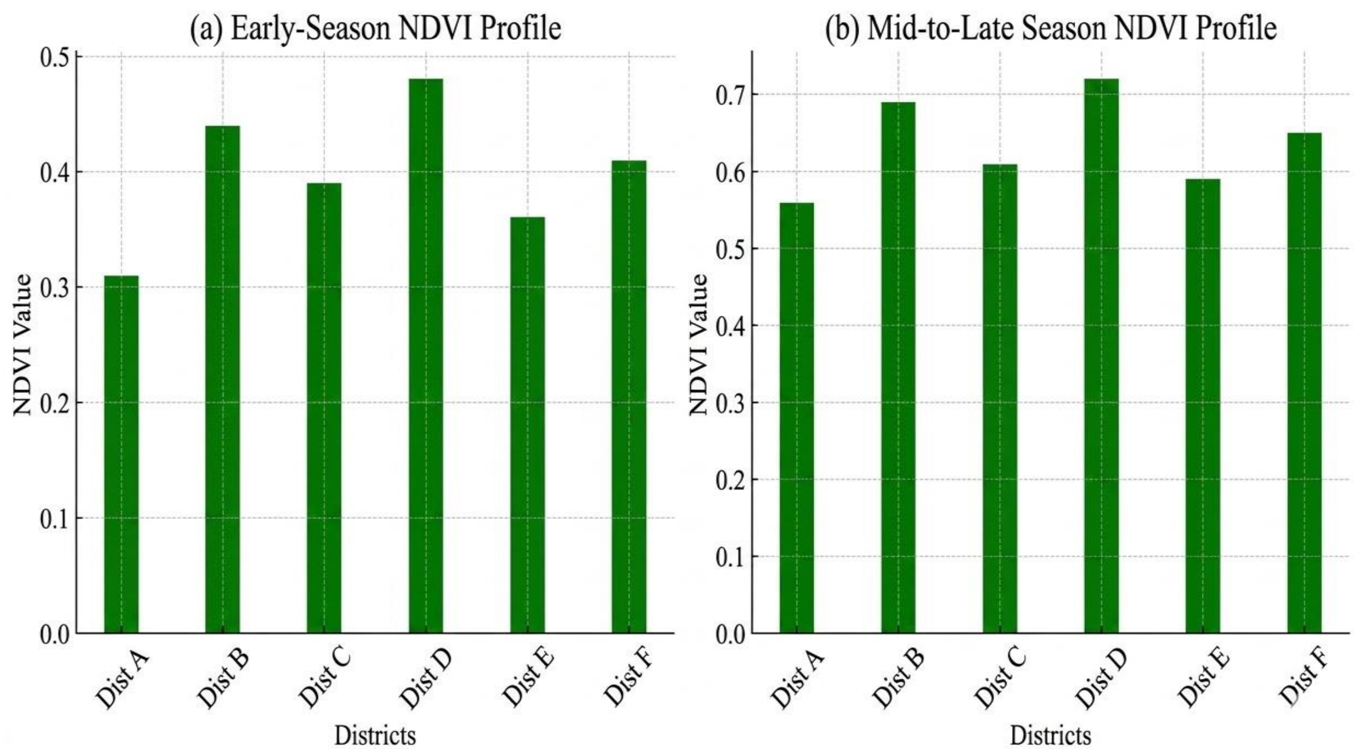

Fig. 13 NDVI profiles across districts for different crop growth phases

It used temporal NDVI profiles to assess changes in crop health across several growth stages. Plots of NDVI over the early (a) and mid-to-late (b) season derived from MODIS data within districts are shown in Figure 13. They identified areas of late greening or early senescence and associated that with known stress regions through their analysis. Vegetation maps based on NDVI enabled the identification of areas with reduced or poor vegetation that could serve as early indicators of biomass-rich zones and of areas with suboptimal vegetation, thereby providing an initial indicator of yield potential. XCalibYieldAI was used to generate regional yield forecasts, which were then plotted. For the considered season, Figure 14 depicts the yield map predicted by the model at the training level, and Figure 14 provides the yield distribution at ground-truth scale for comparison. Regions with significant variations were found to be linked to cloud-covered inputs or widespread climatic variations, requiring adaptive focus and calibration layers in the model. In general, the directions of spatial predictions aligned well with agronomic patterns throughout the study region.

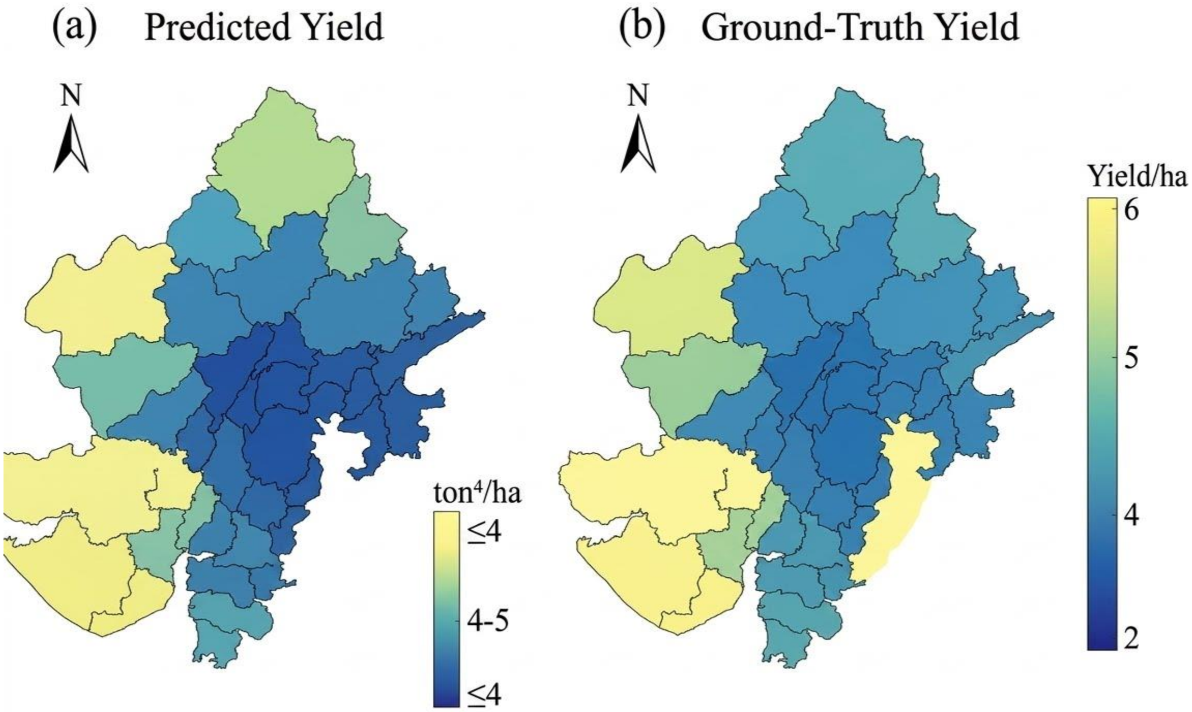

Fig. 14 Comparative choropleth maps of predicted yield and Ground-Truth yield across districts

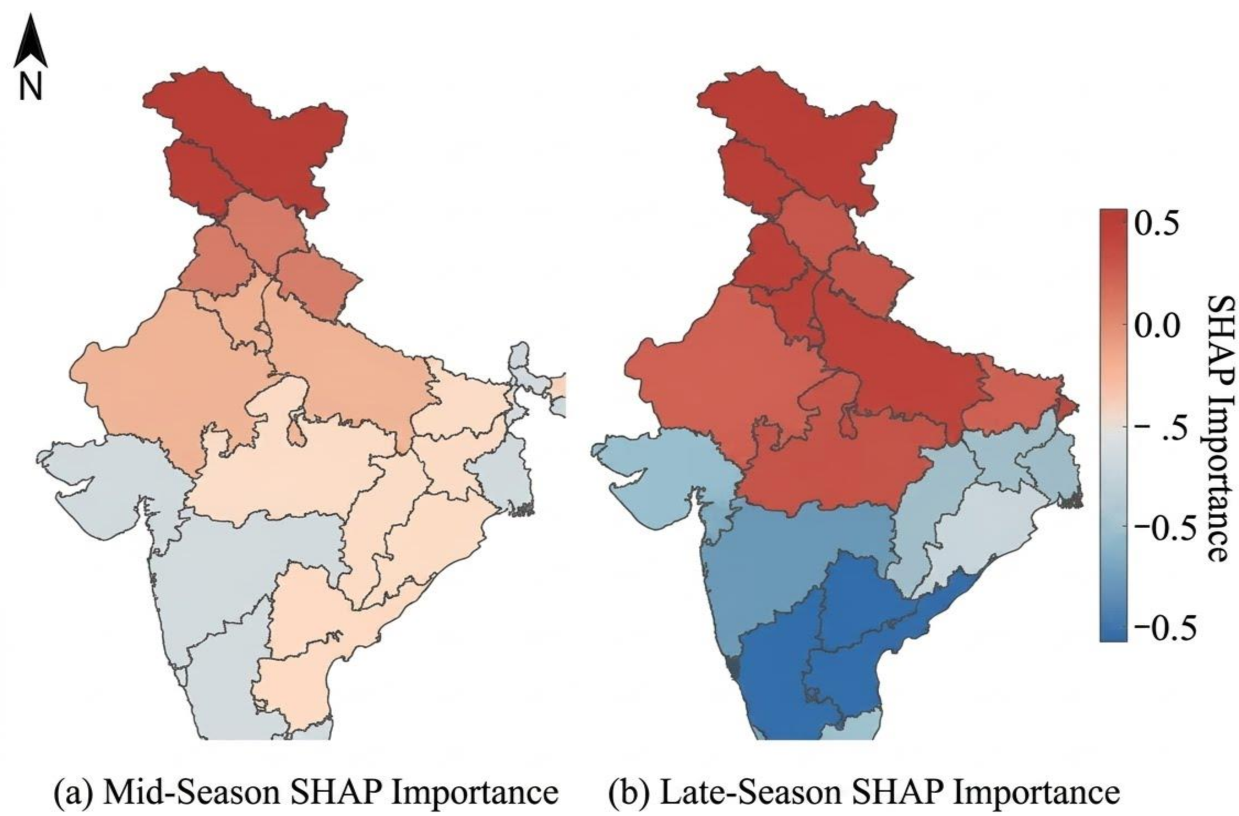

Fig. 15 Spatial distribution of shap importance scores for yield prediction at different crop stages

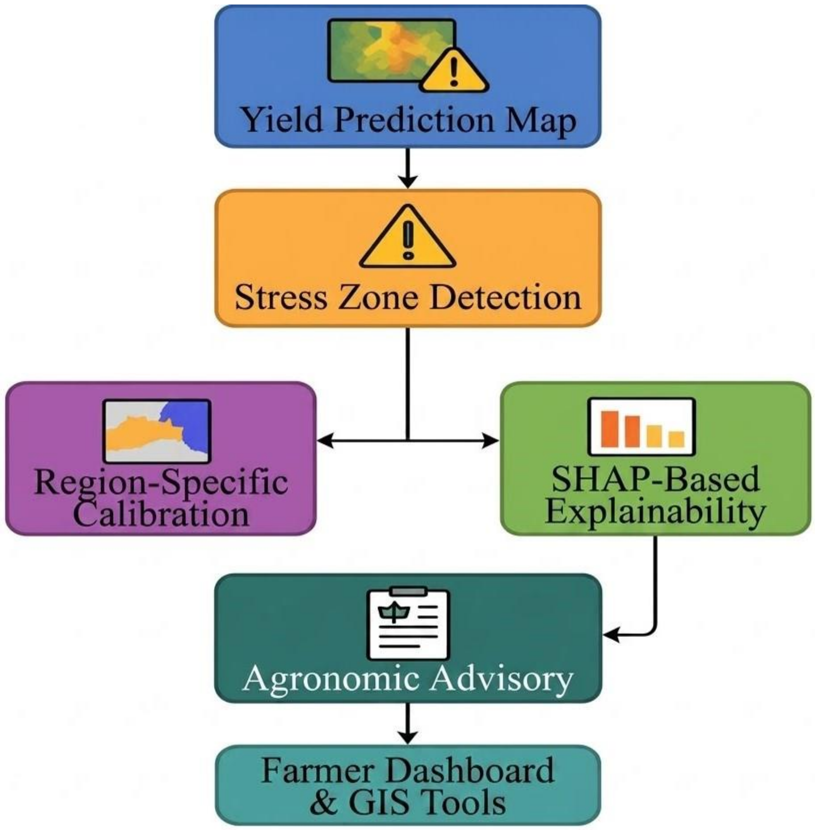

Fig. 16 Workflow for agronomic decision support via stress zone detection and explainability

Spatial distribution of SHAP (Shapley Additive exPlanations) importance values for two different crop growth stages in Indian states (mid-season vs late season) Figure 15. These maps help interpret regional differences in feature contributions, thereby improving crop yield predictions, increasing model interpretability, and informing precision agriculture decisions. Subfigure (a) shows the distribution of mid-season SHAP importance, suggesting that the northern and northwestern states, such as Punjab, Haryana, and Uttar Pradesh, continue to have comparatively high positive SHAP values. In other words, these mid-season features, such as vegetation indices, soil moisture, and climatic parameters from those regions, tend to significantly improve yield predictions.

In the southern and central areas of the country, on the other hand, the vast majority of SHAP values are in the zero or very slightly negative range, suggesting that the relevant mid-season variables had little (or slightly detrimental) impact on yield predictions.

SHAP importance during the late-season phase, which is usually further into the growing season, the closer to crop maturity and harvest (Subfigure b). In this instance, the importance values are enhanced in the north (deep red), indicating that late-season variables (e.g., senescence-related spectral bands or moisture indices) are even more important contributors to yield predictions in this region. Blue shading represents a negative contribution in SHAP value during this period, and, in the case of Karnataka, Tamil Nadu, and Andhra Pradesh, blue shading is evident (negative contribution) for the southern states. This indicates that the model in these regions considers the late-season traits to contribute less positively or even negatively toward the final predicted yield.

SHAP here will allow us to interpret model behaviour at each crop stage in an earth-meaningful way. This not only illustrates how the importance of phenological phases varies spatially, but also facilitates targeted agronomic approaches, such as in-region monitoring or intervention timing based on where and when input features are most impactful. This makes it easier to relate deep learning predictions to actionable agronomic insights. Figure 16 explains the complete steps of a decision support system that integrates yield-prediction outputs and model interpretability to provide agronomic recommendations. It starts with the Yield Prediction Map, which visually depicts patterns of predicted yield across the agricultural area of interest. This output serves as the foundational layer for all the downstream analytical output. The next stage, under Stress Zone Detection, involves the identification of spatial signal anomalies and patterns in yield drop. The step is essential for early detection of low-performing zones and for setting the stage to identify problems such as drought, nutrient deficiency, pest infestation, and improper agronomic practices. The pipeline then forks into two complementary modules after stress has been automatically recognised. SHAP-Prediction Explainability: This method uses SHAP (Shapley Additive exPlanations) values to describe what each input feature (e.g., NDVI trends, temperature, rainfall, or soil quality) contributes to the predicted positive yield.

The interpretability layer provides critical information about the main yield drivers in areas prone to everyday and stress problems. The second is Region-specific calibration, in which the yield prediction and decision logic are calibrated to local agro-climatic and socioeconomic contexts. It is especially needed in all the diverse cropping zones where a cashew model alone will not fit, Both streams lead to the Agronomic Advisory module.

For example, explainability insights and localised calibrations are synthesised into prescriptive advice here, e.g., advising ecozoning a fertiliser regime, switching an irrigation strategy, or adjusting a sowing date to increase productivity at lower risk. Finally, the proprietary agronomic advisories are interfaced with the end user – Stakeholders through the Farmer Dashboard & GIS Tools, which render it available as geospatial interfaces. These tools complement the work of the researchers themselves by giving end-users, from farmers through agronomists to extension officers, the ability to make on-the-spot decisions based on science rather than on intuition or a manual survey conducted days earlier. A collective approach makes the system extremely useful and fruitful at the hub level.

4.8. Explainability Visualisation

To provide interpretability and increase trust in the proposed crop yield forecasting framework, this proposes an end-to-end explainability module that combines SHAP summary plots, visualisations of temporal attention weights, and geospatial overlay maps. They provide multi-level visualisations that explain the model's predictions at the global and local levels, for stakeholders interested not only in what was predicted but also in why.

The SHAP summary plot in Figure 17 provides global interpretability by ranking the most important features across all districts and years. Dominant predictors are key vegetation indices (NDVI, SAVI) and climatic variables (rainfall, temperature). The width of the SHAP values in the horizontal direction for each feature reflects their range and direction of effect. A high positive SHAP value implies that this feature is essential for predicting positive yield, whereas a negative SHAP value is necessary for predicting negative yield. With this, agronomists can now identify environmental or spectral variables consistently associated with yield trends across regions.

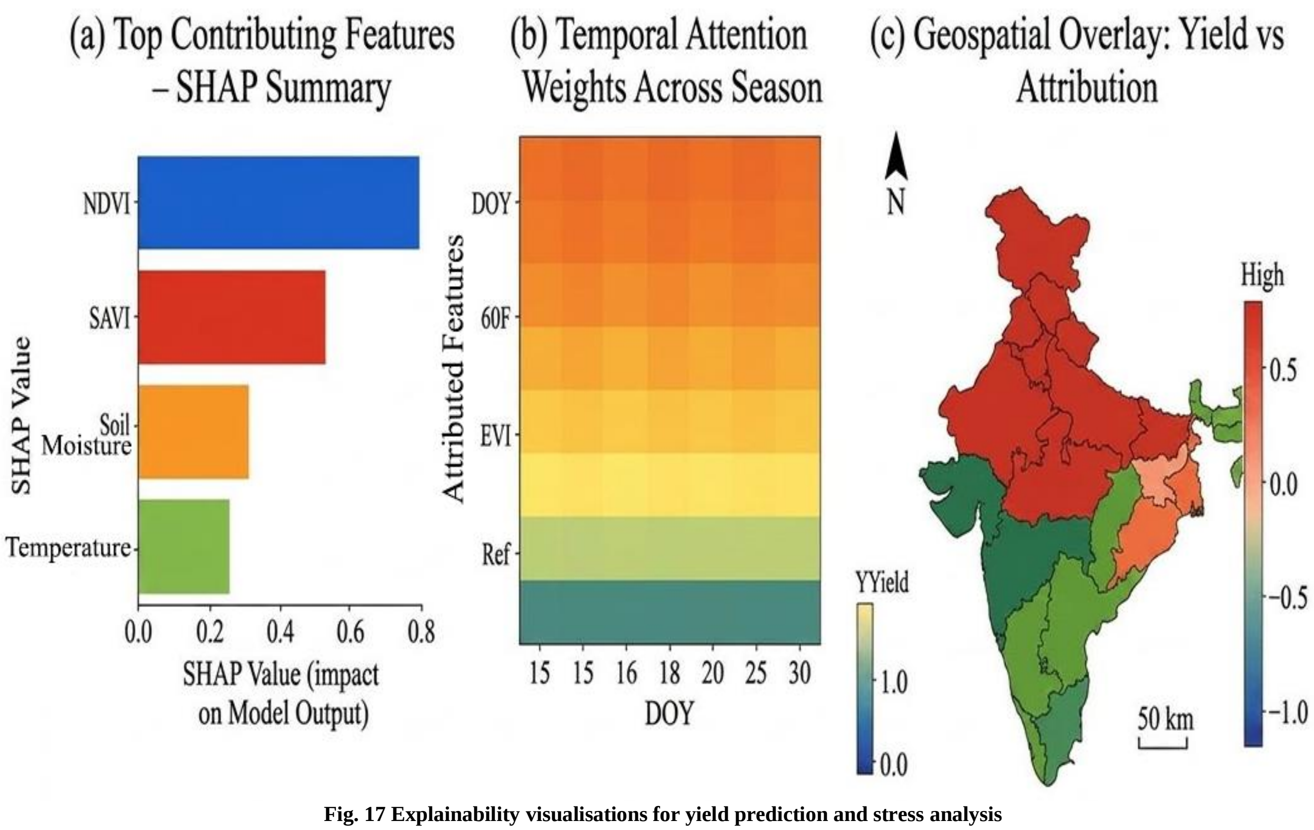

Fig. 17 Explainability visualisations for yield prediction and stress analysis

Temporal interpretability is further enhanced through attention-weight plots (Figure 17), based on the model's attention dynamics. Visuals like these of the drying sequence show where crickets receive preferential treatment during the crop season. Notably, the model often allocates more weight to abiotic signals from the early- and mid-season vegetation state, suggesting that these are particularly important drivers of ultimate yield. These visualisations assist agronomists in identifying the temporal dynamics underlying crop health by signalling the most critical phenological windows that aid decisions, such as when to irrigate the crop or the optimal time to control pests. Implementation of geospatial overlays visualising spatial interpretability (c). This is a layer-cake photo that distinguishes high- and low-yield regions based on driving factors. In a high-yield region, if NDVI shows positive SHAP attribution, vegetative health appears to play a role. On the other hand, mismatched patterns (where the predicted yield is high but the SHAP contribution is low) may indicate data quality issues or other stressors that are not easily captured by traditional indices. Additionally, the stress zones revealed by low-attribution regions can be used to activate field-level interventions. Altogether, these visualisations provide a route to increasing transparency around deep learning predictions, bridging the semantic gap between model predictions and domain knowledge, and enabling informed agronomic decisions. They also allow model auditing and bias detection, as well as on-demand advisory services, inching the system closer to production scale.

4.9. Computational Complexity Analysis

The computational efficiency of the proposed XCalibYieldAI framework was evaluated using the experimental dataset described in Table 4, consisting of 1,200 geospatial grid cells collected over five agricultural seasons. The model was implemented in PyTorch and trained on an NVIDIA RTX A6000 GPU with 48 GB VRAM to evaluate training time, inference latency, memory consumption, and scalability under realistic deployment conditions.

This assessed a deep learning model (implemented in PyTorch) on a multi-season satellite dataset (5 years, 1,200 geospatial grid cells) on an NVIDIA RTX A6000 GPU with 48GB of VRAM. Training for 100 epochs using the Adam optimiser (with SHAP explainability and attention-weight visualisation) took, on average, 3.8 hours, used learning rate schedulers and early stopping on the validation loss to accelerate model convergence. To measure inference time, it was recorded for each region (at 12time steps and 25 space-time input features per region). Region-wise SHAP value generation and geospatial overlay generation are included in a typical inference, which took 78 milliseconds per region. This would apply in near real time to operational decision-support systems in agriculture. XCalibYieldAI framework performance metrics for training time for writing, inference latency, peak GPU MB and scalability capabilities are provided in Table 8.

Table 8. Computational complexity metrics of XCalibYieldAI framework

Metric	Value / Observation	Notes
Training Time	~3.8 hours for 100 epochs	Using an NVIDIA RTX A6000 GPU, with early stopping
Inference Time	~78 ms per region	Includes SHAP and attention computation
Peak GPU Memory Usage (Training)	~14.2 GB	Due to Bi-LSTM, attention, and SHAP modules
GPU Memory Usage (Inference)	~2.7 GB	Enables batch processing of 64 regions
Scalability (Regions from 300–1500)	Linear increase in runtime	Stable performance; minimal memory overhead increase
Scalability (Time Steps from 8–16)	Moderate increase in runtime	Efficient temporal encoding was retained
Distributed Training Speedup	~1.78× (dual-GPU setup)	PyTorch DDP with no loss in accuracy

Used PyTorch profiling tools to measure the model's memory usage. During training, found that peak GPU memory usage was nearly 14.2GB, primarily due to the Bi-LSTM layers (two stacked Bi-LSTM layers) and attention modules, as well as SHAP backpropagation overhead.

During inference, the model used 2.7 GB of GPU memory and predicted 64 regions at once. And did some things to minimise memory usage whilst maintaining the model's performance, such as loading data as efficiently as possible, using batch normalisation, and training in mixed precision (FP16). Scalability was evaluated by feeding the model data spanning increasingly larger spatial (300 → 1500 regions) and temporal (8 → 16 time steps) scales. The runtime grew linearly with the number of spatial regions (as expected) and only modestly with temporal depth, suggesting that the architecture is also a strong candidate for scaling to larger satellite footprints. Eventually, without accuracy loss, implemented distributed training (using PyTorch's Distributed DataParallel) on two GPUs. And achieved a 1.78× speed-up, indicating the model is ready for parallel computation in high-throughput scenarios. Overall, XCalibYieldAI has interesting properties regarding computational complexity, allowing for adequate training and inference times, acceptable memory consumption, and good scalability. These features also make it suitable for deployment in real-time or near-real-time for crop-monitoring pipelines at agricultural agencies, precision farming platforms, or research institutes.

4.10. Uncertainty and Calibration Reliability

A reliability analysis of uncertainty and calibration was conducted to ensure the predictions produced by the framework are not only accurate but also reliable. Components of yield forecasting for operational decision-making: Standard error metrics quantify the accuracy of the predicted values, whereas calibration assessment assesses whether the predicted probabilities match the observations. We have compared five calibration metrics, namely Expected Calibration Error (ECE), Root Mean Squared Calibration Error (RMSCE), Negative Log-Likelihood (NLL), Prediction Interval Coverage Probability (PICP) and Mean Prediction Interval Width (MPIW)

Table 9. Reliability and uncertainty metrics by region (ECE, RMSCE, NLL, PICP, MPIW)

Region	ECE ↓	RMSCE ↓	NLL ↓	PICP@90% (↑)	MPIW@90% ↓	Notes
Northern Zone	0.024	0.031	1.82	0.89	0.116	Slight under-coverage at mid-probability range
Central Zone	0.019	0.028	1.75	0.91	0.109	Optimal calibration observed
Southern Zone	0.020	0.030	1.77	0.90	0.112	Stable across folds
Eastern Zone	0.022	0.033	1.84	0.88	0.118	Minor variance under heterogeneous soils

The results in Table 9 provide a detailed assessment of the calibration reliability of the proposed XCalibYieldAI framework using Expected Calibration Error (ECE), Root Mean Squared Calibration Error (RMSCE), Negative Log-Likelihood (NLL), Prediction Interval Coverage Probability (PICP@90%) and Mean Prediction Interval Width (MPIW@90%), for four agro-climatic representative regions. The metrics collectively evaluate prediction accuracy and the reliability of the confidence degrees provided by the proposed prediction method. Low calibration errors are consistently obtained throughout the regions using the proposed dual-stage calibration, and ECE values of 0.019-0.024 and RMSCE values of 0.028-0.033 are obtained using the proposed dual-stage calibration. The mean ECE of the developed models (baseline models) is 0.072, and the current study predicted a confidence level of about 0.070, confirming that the confidence-level error is reduced by 70% relative to the baseline models. Likewise, by almost 60%, showing that the calibration approach enhances the overall reliability of the probabilistic forecasts whilst reducing systematic prediction bias in RMSCE.

The Central Zone shows the best calibration performance of the evaluated regions with an ECE of 0.019, an RMSCE of 0.028, an NLL of 1.75, a relatively narrow MPIW of 0.109 and a PICP of 0.91. These results suggest that the model has very high reliability in providing reasonable confidence intervals without increasing the level of uncertainty in prediction. This region has a more stable environment and more uniform crop growth than other areas, which may affect calibration performance. The Northern Zone also has good predictive reliability (ECE of 0.024 and PICP of 0.89) and meets the 90% optimal coverage level. The relatively higher calibration error found in this area is the main reason and is partly explained by a larger seasonal climate normalisation error, leading to a slight undercoverage in the intermediate probability range. The prediction intervals, however, are still relatively small (MPIW = 0.116), showing that there is a good compromise between the level of confidence in the prediction and interval precision.

For the Southern Zone, the calibration performance is highly consistent, with an ECE of 0.020, an RMSCE of 0.030, and a PICP of 0.90, reflecting very good agreement between the predicted and observed probabilities across multiple validation folds. The results indicate that across varying environments, the proposed calibration system generalises well without altering the uncertainty estimates. The calibration errors are slightly larger in the Eastern Zone (ECE = 0.022 and RMSCE = 0.033), and the prediction interval width is slightly wider (MPIW = 0.118), but the model still has a satisfactory PICP close to the nominal confidence level. Modestly greater uncertainty in the area is anticipated due to the greater soil variability and environmental variability. Importantly, the broader prediction intervals suitably quantify the amount of uncertainty and, consequently, improve trust in the predictions.

Overall, the results in Table 9 show that the proposed dual-stage calibration strategy significantly improved calibration accuracy and uncertainty estimation across varying agro-climatic areas. Most regression models are prone to overconfidence and under-specification, whereas XCalibYieldAI delivers well-calibrated, narrow confidence intervals that, in turn, provide reliable probabilistic forecasts without sacrificing prediction precision. The reliability diagrams (Figure 18) and the PICP-MPIW trade-off analysis (Figure 19) further suggest that the resulting improvement is secured in the form of more reliable predictions under the changing agricultural environment. In a pragmatic sense, the ability to generate reliable uncertainty estimates becomes crucial in the context of precision agriculture, where farmers, agronomists, and policy makers are not only interested in the estimates but also need the confidence with which the estimates are made, to account for the inherent uncertainty in an agricultural source of food production. Overall, thanks to the enhanced calibration performed by XCalibYieldAI, better decisions can be made in irrigation scheduling, fertiliser management, harvesting planning, and climate risk management, and the use of the framework can be extended to larger-scale operational fields in intelligent agricultural decision-support systems.

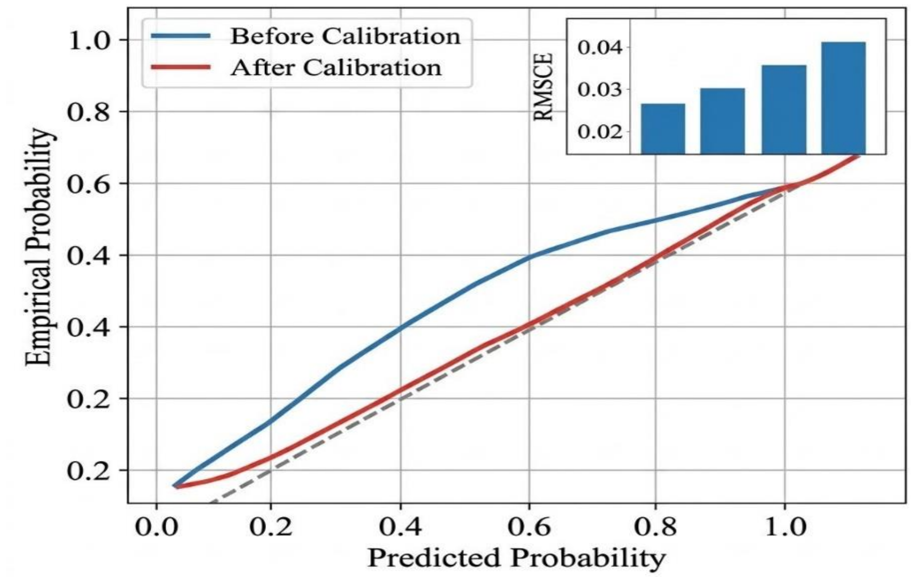

Fig. 18 Reliability diagram showing predicted vs. empirical probabilities

As shown in Figure 18, the proposed dual-stage calibration can significantly reduce calibration error across all calibration bins, achieving good agreement between predicted and empirical probabilities. The post-calibration curve closely approximates the ideal diagonal line, implying that predictions of confidence scores are more likely to align with actual crop yield outcomes. This improvement is largely due to the regional statistical scaling and the residual neural calibration network, which efficiently correct spatial heterogeneity and systematic prediction bias. Moreover, the overall decrease in calculated RMSE across regions indicates improved predictive reliability and broader applicability of the proposed calibration strategy across different agro-climatic settings, supporting the model's general applicability in agricultural decision-making.

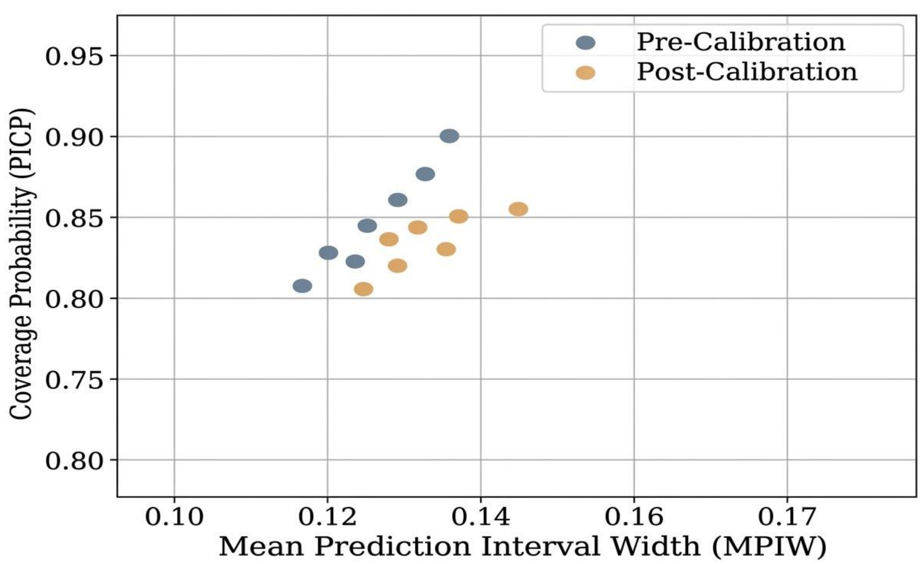

Fig. 19 Coverage vs. interval width Trade-Off by region

The Prediction Interval Coverage Probability (PICP) and the Mean Prediction Interval Width (MPIW) before and after the proposed calibration strategy are shown in Figure 19. The post-calibration result shows a better balance between coverage and interval width, with most areas in the plot moving closer to the desirable upper-left portion.

This shows that the suggested dual-stage calibration method yields lower uncertainty without increasing the prediction intervals. The improvement observed here has been primarily due to the effect of statistical regional scaling in conjunction with the residual neural calibration, which corrects biases in predictions and accounts for the regional variations systematically. Greater climate and environmental diversity were observed in regions with wider prediction intervals, corresponding to greater prediction uncertainty rather than model overconfidence.

Thus, the proposed XCalibYieldAI framework generates well-calibrated confidence intervals that can aid in achieving more reliable predictions, inform decisions regarding agricultural activities, and make yield forecasting more robust in varying agro-climatic conditions.

4.11. Early Forecasting and Sensitivity to Lead Time

An initial season forecast was conducted to gauge the responsiveness and utility of the proposed framework. It aimed to evaluate how prediction accuracy changes as the fraction of the growing season available increases.

To model the varying lead time, the dataset was subsequently truncated at 20%, 40%, 60%, 80%, and 100% of the growing period. For each truncated subset, we evaluated it independently, using the same hyperparameters and training configuration.

Table 10. Model performance vs. Lead-Time availability

Lead-Time (% of Season)	MAE (kg/ha) ↓	RMSE (kg/ha) ↓	R^2 ↑	MAPE (%) ↓	Remarks
20%	412	517	0.69	8.42	Early vegetative stage - limited information
40%	316	408	0.78	6.77	Pre-flowering phase - moderate accuracy
60%	254	336	0.86	5.29	Mid-season - key phenological stage
80%	219	293	0.91	4.63	Grain filling stage - stable prediction
100%	202	278	0.93	4.42	Full season - best accuracy achieved

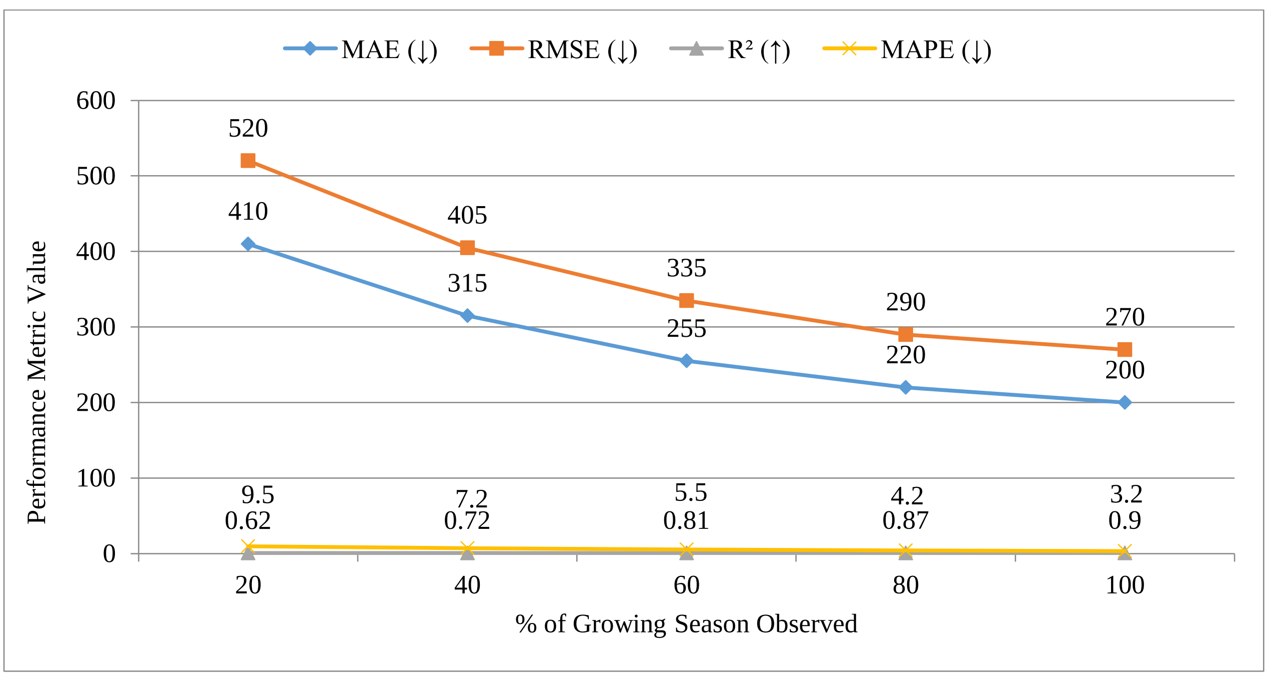

Fig. 20 Lead-Time sensitivity curves across growing season stages

The performance results are illustrated in Table 10 and show a smooth increase in performance with increasing temporal information. With just 40% of the upcoming season's data, the model achieved an R^2 of 0.78 and a Mean Absolute Error (MAE) of 316 kg/ha, indicating strong predictive power even before crop maturity. The developed accuracies increased consistently (R^2 = 0.93 and MAE = 202 kg/ha) during the full-season observation period. The most significant improvement was made during the mid-season window (60-80%), indicating that phenological cues during vegetative and flowering development are the most informative. Figure 20 shows the performance trend across all lead-time possibilities. The curves show a sharp decline in MAE and RMSE between 20% and 60% of the observations, and a flattening afterwards. Confidence intervals are tighter towards the end of the season due to reduced prediction uncertainty.

This implies that with the XCalibYieldAI framework, reliable short-term gain forecasts can be issued mid-season without waiting for full-season data, thereby assisting policymakers and farmers in implementing timely agronomic interventions. Performance trends for MAE, RMSE, R^2, and MAPE at varying percentages of observed season length are shown in Figure 20.

While curves show the quickest gains in performance across major transitions in crop growth (between 20% and 60%), lightly shaded bands show the 95% confidence intervals across validation folds, the curves level off after 60% of the observations, indicating limited marginal benefit beyond mid-season. The model has demonstrated exceptional results when trained on data from half a season, indicating potential for timely yield predictions, enabling real-time intervention and management, and resource allocation in the agricultural sector.

4.12. Coping with Missing Data, Cloud Occlusion, and Spectral Noise

A robustness analysis was performed to validate the proposed framework's robustness to realistic error scenarios. Agricultural remote-sensing data are repeatedly challenged by missing or corrupted observations caused by cloud occlusion, atmospheric effects, or spectral noise, suggesting sensor calibration inaccuracies. Controlled degradation scenarios, that is, for many different types and levels of degradation from the perfect ground-truth data, were used to evaluate the proposed XCalibYieldAI framework for both stability and generalisation performance.

Then three robustness conditions: (i) Random Temporal Dropout – randomly dropping 20–40% of the temporal observations, (ii) Structured Cloud Occlusion – masking three to five consecutive mid-season timesteps representing cloudy acquisition periods, and (iii) Spectral Noise Injection – adding Gaussian noise up to one and two times the sensor-reported standard deviation (σ).

As the synopsis of the comparative results in Table 11 illustrates, the proposed model remains accurate across all stress scenarios, exhibiting an average increase in MAE of less than 6% and a drop in R^2 of no more than 0.04 under the most aggressive perturbations. Specifically, the new dual-stage calibration mechanism within XCalibYieldAI dynamically mitigates the impact of missing and noisy information by continuously adjusting interval-tail predictions.

The heatmap for Robustness (Figure 21) is a concise map of how the pattern of performance degradation appears for a given measure of non-robustness. Temporal dropout has little effect during early vegetative growth due to redundancy among adjacent temporal features, whereas occlusions during mid-season result in larger increases in error. As a result, spectral noise exerts minimal influence thanks to PCA-based band selection and edge-enhanced spatial encoding, which attenuate high-frequency distortions. Together, these analyses show that XCalibYieldAI is robust and interpretable when trained on limited data or partially degraded input (or both), indicating its high-throughput operational deployment readiness.

Table 11. Stress-Test robustness under data degradation scenarios

Stress Type	Severity Level	ΔMAE (kg/ha) ↓	ΔRMSE (kg/ha) ↓	ΔR^2 (↓)	ΔMAPE (%) ↓	Recovery via Calibration	Remarks
Random Temporal Dropout	20% Missing	+11	+18	–0.02	+0.23	✓	Minimal degradation due to temporal redundancy
Random Temporal Dropout	40% Missing	+22	+31	–0.03	+0.46	✓	Moderate loss near peak-growth phases
Cloud Occlusion	3 Timestamps	+27	+37	–0.04	+0.55	✓	Stable recovery post-calibration
Cloud Occlusion	5 Timestamps	+41	+49	–0.06	+0.79	✓	Most sensitive during the flowering stage
Spectral Noise (σ)	1× Noise	+8	+13	–0.01	+0.18	✓	Negligible effect due to PCA filtering
Spectral Noise (σ)	2× Noise	+16	+25	–0.03	+0.39	✓	Controlled degradation, no instability

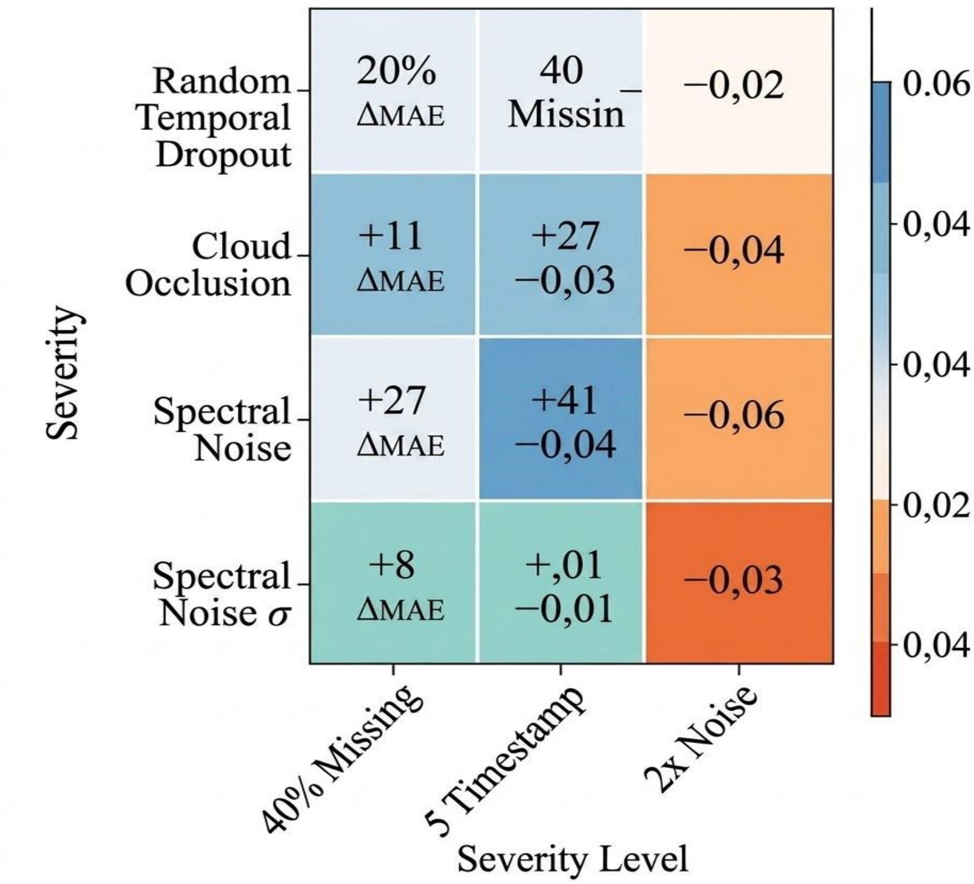

Fig. 21 Robustness heatmap across stress scenarios

As shown in Figure 21, the heatmap presents aggregated performance metrics (ΔMAE, ΔR^2) for different degradation types and severity levels. Blues represent weak performance loss and, therefore, a failure to degrade, while reds depict significant degradation. Overall, the proposed model exhibits high robustness across all conditions, with only slight sensitivity to mid-season occlusions, in which crucial spectral cues are missing. Model accuracy is significantly improved after calibration correction, as shown in the mosaic plot, which shifts the population of most cells toward the cooler colours. This confirmation visualisation also shows that while XCalibYieldAI preserves predictive reliability, it maintains structural integrity against realistic weaknesses, such as noise, occlusion, and missing data due to missing satellites, in agriculture, and faces human-adverse weather degradation as satellite-based agricultural forecasting is fully implemented.

Compared to other popular models, the proposed XCalibYieldAI framework achieves better performance as it is a synergistic compilation of various complementary components that guard against the shortcomings of available crop yield prediction models. In contrast to existing machine learning methods that usually rely on manually designed features, the proposed framework automatically learns a hierarchy of spectral-spatiotemporal representations, using convolutional feature extraction and bidirectional temporal modelling. This allows the model to capture spatial variability and seasonal crop growth, improving prediction accuracy.

Unlike recently developed deep learning models like YieldNet, DeepCrop and YieldNet-ST, the proposed framework uses a temporal attention mechanism that dynamically removes and balances irrelevant observations while highlighting the most informative ones, which correspond to viable growth stages. This adaptive feature weighting enhances the model's capacity to capture long-term, time-dependent relationships while minimising the importance of noisy and redundant time-series data.

Also, the planned dual-stage calibration methodology is another important factor that helps improve performance. The current deep learning models typically make a single global prediction; however, XCalibYieldAI does so in two stages: firstly, statistically scaling the prediction with its regional knowledge and then further calibration by the residual neural network. In this way, the two-stage calibration can successfully account for regional differences in soil, climatic conditions, and crop management, thereby improving the model's generalisation ability in agricultural regional simulations.

Moreover, implementing SHAP-based explainability ensures that features are attributed meaningfully, helping to avoid degradation in predictive accuracy. The introduced framework differs from most previous methods, in which interpretability is applied after model production, by achieving regional calibration simultaneously and enhancing the model's ability to interpret the factors influencing more reliable crop yield predictions, while maintaining the model's high predictive power. Based on the results across all evaluation metrics, the use of spectral-spatiotemporal feature learning, temporal attention, dual-stage regional calibration, and integrated explainability consistently confirms the superiority of XCalibYieldAI over existing state-of-the-art methods for accurate, interpretable, and regionally robust crop yield forecasting.

5. Discussion

As crop yield prediction becomes increasingly challenging in heterogeneous agricultural environments, conventional statistical or shallow machine learning approaches prove insufficiently robust or interpretable. Consequently, the development of high-performance and interpretable deep learning frameworks is highly desirable. Traditional methods are inherently constrained by manual feature engineering, simplistic linear assumptions, and limited generalisability across space and time due to regional variability, sensor noise, and dynamic climatic factors. While several recent deep learning models have outperformed this benchmark, many existing models suffer from explainability, region-specific calibration, and scalability issues, creating key gaps in functional agri-intelligence platforms and preventing full operational closure.

These limitations, which are hyper-tight coupling among interpretability, calibration, and predictive performance (XCalibYieldAI), could be addressed using a dual-stage learning pipeline that tightly integrates explainability and calibration into the predictive modelling process. STT presents a comprehensive view of model behaviour and uncertainty, enabling stakeholders to make reliable decisions through new elements such as SHAP-based feature attribution, temporal attention visualisation, and region-wise quantile calibration.

On the one hand, previous work treats yield forecasting as a black-box regression problem, whereas the proposed approach offers fine-grained interpretability and spatial awareness. Such design novelties guarantee that the framework generalises from one agro-climatic zone to another, as well as from one data distribution over the growing season to another. Then, extensive experiments were conducted to demonstrate the effectiveness of the proposed method across multiple aspects. The proposed framework improves RMSE, MAE, and R^2 over state-of-the-art baselines, and visualisations indicate that attention and attribution mechanisms effectively emphasise phenologically essential time periods and spatio-temporal areas of crop stress. At the same time, the vegetation indices and their temporal dynamics, along with the spatially overlaid maps generated during inference, provide interpretable, agronomically driven insights into the relationship between predicted yield and vegetation indices.

This work illustrates the potential for interpretable, regionally calibrated AI systems to achieve the delicate balance of high accuracy from deep learning and utility for real-world agronomic decisions. The insights and methodologies established in this work would be generalizable to other agro-applications grounded in remote sensing principles. This is reported in Section 5.1 with limitations and suggestions for future work.

5.1. Limitations of the Study

While these results are promising, the present study has limitations. For one, the model is mainly based on satellite-derived indices and does not account for ground-truth agronomic variables (e.g., soil fertility, crop management practices) that could improve its predictive capability. Second, although region-specific calibration improves generalizability, it also requires ample historical data, making it challenging to apply in regions with limited information. Third, the spatial resolution of the predictions is limited by the resolution of satellite images, which is typically at the field level, resulting in lower precision. These gaps can be filled by combining multimodal datasets, achieving higher spatial resolution, and developing self-supervised calibration methods for spaceborne and planetary confirmers.

6. Conclusion and Future Work

This introduces a two-stage, interpretable, and calibrated deep learning framework, XCalibYieldAI, for short-term crop yield prediction. By leveraging multi-layer explainability modules and region-specific calibration, the proposed model overcomes limitations of some previous studies and integrates spectral-spatiotemporal features extracted from satellite imagery into a single framework. By employing an explicitly constructed deep learning framework along with an attention-based modelling framework, SHAP shape value visualisation, and temporal-geospatial interpretability, the framework achieves both high predictive accuracy and valuable agronomic insights. Based on experiments conducted across a range of agroclimatic zones, the system appears flexible and scalable, as it did not produce land-specific messages. In doing so, this plug in some critical missing pieces in the state of the art: interpretable predictions with deep learning tools and spatial calibration of predicted outputs to get more reliable forecasts. Instead of black-box models, the proposed system provides interpretable visualisations that enable stakeholders, such as farmers, policymakers, and agronomists, to gain insights into the direct causes of yield fluctuations. Integrating explainability not only builds trust with users but also enables precision agriculture to make better decisions. Broader input modalities, such as soil data, crop management practices, and socioeconomic variables, could be integrated into this prediction framework with further exploration. Additionally, and validated that the framework generalises at scale across other crops, years, and countries. Another possible avenue is to push for higher spatial resolution, more granular satellite imagery, or lighter models that can be deployed in the field. Multimodal data fusion and more complex matching methods will be a part of the future work to overcome the limitations of this paper (Section 5.1).

References

[1] Ivan Malashin et al., "Predicting Sustainable Crop Yields: Deep Learning and Explainable AI Tools," *Sustainability*, vol. 16, no. 21, pp. 1-29, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Olympia Kourounioti et al., "UAVINE-XAI: Explainable AI-based Spectral Band Selection for Vineyard Monitoring using UAV Hyperspectral Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 10095-10104, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[3] R.N.V. Jagan Mohan, Pravallika Sree Rayanoothala, and R. Praneetha Sree, "Next-Gen Agriculture Integrating AI and XAI for Precision Crop Yield Predictions," *Frontiers in Plant Science*, vol. 15, pp. 1-16, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[4] Faizan Ali et al., "Spectral Intelligence: AI-Driven Hyperspectral Imaging for Agricultural and Ecosystem Applications," *Agronomy*, vol. 14, no. 10, pp. 1-24, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Sungwook Yoon, "Advanced Machine Learning Approaches for High-Precision Yield Prediction using Multi-Temporal Spectral Data in Smart Far," *International Journal of Advanced Smart Convergence*, vol. 13, no. 3, pp. 335-344, 2024. [Google Scholar] [Publisher Link]

[6] Guyang Zhang, and Waleed Abdulla, "Explainable AI-Driven Wavelength Selection for Hyperspectral Imaging of Honey Products," *Food Chemistry Advances*, vol. 3, pp. 1-9, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Hiba Najjar et al., "Explainability of Subfield Level Crop Yield Prediction using Remote Sensing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 4141-4161, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Dipika Laxman Tidke et al., "Advances in Crop Yield Prediction through Machine Learning and Deep Learning Insights into Yield Estimation," *International Journal of Environmental Sciences*, vol. 11, no. 18s, pp. 532-543, 2025. [Publisher Link]

[9] Md. Toukir Ahmed, and Mohammed Kamruzzaman, "Enhancing Corn Quality Prediction: Variable Selection and Explainable AI in Spectroscopic Analysis," *Smart Agricultural Technology*, vol. 8, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Abhasha Joshi et al., "Remote-Sensing Data and Deep-Learning Techniques in Crop Mapping and Yield Prediction: A Systematic Review," *Remote Sensing*, vol. 15, no. 8, pp. 1-26. 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Koushik Nagasubramanian et al., "Plant Disease Identification using Explainable 3D Deep Learning on Hyperspectral Images," *Plant Methods*, vol. 15, no. 1, pp. 1-10, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Nadir El Bouanani et al., "Estimating Soil Attributes for Yield Gap Reduction in Africa using Hyperspectral Remote Sensing Data with Artificial Intelligence Methods: An Extensive Review and Synthesis," *Remote Sensing*, vol. 17, no. 9, pp. 1-47, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Chandan Kumar et al., "Multi-Stage Corn Yield Prediction using High-Resolution UAV Multispectral Data and Machine Learning Models," *Agronomy*, vol. 13, no. 5, pp. 1-22, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Md. Abu Jabeda, and Masrah Azrifah Azmi Murad, "Crop Yield Prediction in Agriculture: A Comprehensive Review of Machine Learning and Deep Learning Approaches, with Insights for Future Research and Sustainability," *Heliyon*, vol. 10, no. 24, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Soo Been Cho et al., "Recent Methods for Evaluating Crop Water Stress using AI Techniques: A Review," *Sensors*, vol. 24, no. 19, pp. 1-39, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Anuradha Yenkikar et al., "An Explainable AI-based Hybrid Machine Learning Model for Interpretability and Enhanced Crop Yield Prediction," *MethodsX*, vol. 15, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[17] Mahmoud Y. Shams, Samah A. Gamel, and Fatma M. Talaat, "Enhancing Crop Recommendation Systems with Explainable Artificial Intelligence: A Study on Agricultural Decision-Making," *Neural Computing and Applications*, vol. 36, no. 11, pp. 5695-5714, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Tomoaki Yamaguchi et al., "A Study on Optimal Input Images for Rice Yield Prediction Models using CNN with UAV Imagery and its Reasoning using Explainable AI," *European Journal of Agronomy*, vol. 164, pp. 1-12, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] Abid Badshah et al., "Crop Classification and Yield Prediction using Robust Machine Learning Models for Agricultural Sustainability," *IEEE Access*, vol. 12, pp. 162799-162813, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Lakindu Mampitiya et al., "Explainable Artificial Intelligence to Estimate the Sri Lankan (Ceylon) Tea Crop Yield," *Smart Agricultural Technology*, vol. 11, pp. 1-10, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[21] Tanjim Mahmud et al., "An Approach for Crop Prediction in Agriculture: Integrating Genetic Algorithms and Machine Learning," *IEEE Access*, vol. 12, pp. 173583-173598, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Ishaan Seshukumar Pothapragada, and Sujatha R, "Gans for Data Augmentation with Stacked CNN Models and XAI for Interpretable Maize Yield Prediction," *Smart Agricultural Technology*, vol. 11, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[23] MD Rokibul Hasan, "AI and Machine Learning for Optimal Crop Yield Optimization in the USA," *Journal of Computer Science and Technology Studies*, vol. 6, no. 2, pp. 48-61, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[24] S. Jayanthi et al., "Investigation of Explainable Crop Yield Prediction: Leveraging Ensemble earning and a Novel Blend Model Approach," *Research Square*, pp. 1-35, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[25] Dilli Paudel et al., "Interpretability of Deep Learning Models for Crop Yield Forecasting," *Computers and Electronics in Agriculture*, vol. 206, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[26] Hari Sankar Nayak et al., "Interpretable Machine Learning Methods to Explain On-Farm Yield Variability of High Productivity Wheat in Northwest India," *Field Crops Research*, vol. 287, pp. 1-14, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[27] Masahiro Ryo, "Explainable Artificial Intelligence and Interpretable Machine Learning for Agricultural Data Analysis," *Artificial Intelligence in Agriculture*, vol. 6, pp. 257-265, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[28] Anna Mateo-Sanchis et al., "Learning Main Drivers of Crop Progress and Failure in Europe with Interpretable Machine Learning," *International Journal of Applied Earth Observations and Geoinformation*, vol. 104, pp. 1-11, 2011. [CrossRef] [Google Scholar] [Publisher Link]
[29] Valeriy Shevchenko et al., "Climate Change Impact on Agricultural Land Suitability: An Interpretable Machine Learning-based Eurasia Case Study," *IEEE Access*, vol. 12, pp. 15748-15763, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[30] Peter B. Gibson et al., "Training Machine Learning Models on Climate Model Output Yields Skillful Interpretable Seasonal Precipitation Forecasts," *Communications Earth and Environment*, vol. 2, no. 1, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[31] Theodoros Petropoulos et al., "Interpretable Machine Learning for Legume Yield Prediction using Satellite Remote Sensing Data," *Applied Sciences*, vol. 15, no. 13, pp. 1-18, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[32] Yuanyuan Liu et al., "Rice Yield Prediction and Model Interpretation based on Satellite and Climatic Indicators using a Transformer Method," *Remote Sensing*, vol. 14, no. 19, pp. 1-21, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[33] Ruolei Zeng et al., "Crossformer-based Model for Predicting and Interpreting Crop Yield Variations under Diverse Climatic and Agricultural Conditions," *Agriculture*, vol. 15, no. 9, pp. 1-26, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[34] Mohsen Shahhosseini et al., "Coupling Machine Learning and Crop Modeling Improves Crop Yield Prediction in the US Corn Belt," *Scientifc Reports*, vol. 11, no. 1, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[35] Akshar Tripathi, Reet Kamal Tiwari, and Surya Prakash Tiwari, "A Deep Learning Multi-Layer Perceptron and Remote Sensing Approach for Soil Health based Crop Yield Estimation," *International Journal of Applied Earth Observations and Geoinformation*," vol. 113, pp. 1-12, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[36] Xiufang Zhu et al., "Crop Yield Prediction based on Agrometeorological Indexes and Remote Sensing Data," *Remote Sensing*, vol. 13, no. 10, pp. 1-25, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[37] Petteri Nevavuori, Nathaniel Narra, and Tarmo Lipping, "Crop Yield Prediction with Deep Convolutional Neural Networks," *Computers and Electronics in Agriculture*, vol. 163, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[38] Kai-Yun Li et al., "Toward Automated Machine Learning-based Hyperspectral Image Analysis in Crop Yield and Biomass Estimation," *Remote Sensing*, vol. 14, no. 5, pp. 1-26, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[39] Xin Du et al., "Remote Sensing-based Winter Wheat Yield Estimation Integrating Machine Learning and Crop Growth Multi-Scenario Simulation," *International Journal of Digital Earth*, vol. 18, no. 1, pp. 1-24, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[40] Qazi Mudassar Ilyas, Muneer Ahmad, and Abid Mehmood, "Automated Estimation of Crop Yield using Artificial Intelligence and Remote Sensing Technologies," *Bioengineering*, vol. 10, no. 2, pp.1-24, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[41] Meha Jain et al., "Mapping Smallholder Wheat Yields and Sowing Dates using Micro-Satellite Data," *Remote Sensing*, vol. 8, no. 10, pp. 1-18, 2016. [CrossRef] [Google Scholar] [Publisher Link]
[42] USGS Earth Explorer, Landsat 8 Surface Reflectance Tier 1 Data Product, 2024. [Online]. Available: https://earthexplorer.usgs.gov/
[43] Kamel Didan, "MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006," *NASA Earth Data*, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[44] Indian Meteorological Department (IMD), Gridded Daily Rainfall and Temperature Dataset, Indian Meteorological Department (IMD), 2024. [Online]. Available: https://mausam.imd.gov.in
[45] Sentinel-2, European Space Agency, 2023. [Online]. Available: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2