Discover Artificial Intelligence

https://doi.org/10.1007/s44163-026-01912-x

Article in Press

A critical review of deep learning methods for trustworthy crop yield prediction

Received: 14 April 2026

Accepted: 27 July 2026

Published online: 05 August 2026

Cite this article as: Taremwa D., Ahishakiye E., Obbo A. *et al.* A critical review of deep learning methods for trustworthy crop yield prediction. *Discov Artif Intell* (2026). https://doi.org/10.1007/s44163-026-01912-x

Danison Taremwa, Emmanuel Ahishakiye, Aggrey Obbo, Paul Kategaya Kisozi & Fred Kaggwa

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

A Critical Review of Deep Learning Methods for Trustworthy Crop Yield Prediction

Danison Taremwa*1,5, Emmanuel Ahishakiye2, Aggrey Obbo3, Paul Kategaya Kisozi[4], Fred Kaggwa [5]

[1]Department of Computer Science, Kyambogo University, Kampala, Uganda
[2]Department of Networks, Data Science and Artificial Intelligence, Kyambogo University, Kampala, Uganda
[3]Department of Software Engineering, Mbarara University of Science & Technology, Mbarara, Uganda
[4]Department of Environmental Science, Kyambogo University, Kampala, Uganda
[5]Department of Computer Science, Mbarara University of Science & Technology, Mbarara, Uganda

***Corresponding Author**: Danison Taremwa*; **Email**: taremwa.danison@gmail.com*

Abstract

Deep learning is now central to crop-yield prediction (CYP), but reported gains are too often attributed to architectural novelty rather than to the data regime and evaluation design that actually determine performance. This structured critical review treats CYP as an evidence chain rather than a leaderboard problem. Unlike prior surveys that mainly catalogue architectures or application areas, this review contributes a practical evaluative framework that links architecture choice to data regime, target definition, and validation design, helping practitioners judge model suitability beyond reported accuracy alone. It asks whether each model family matches the structure of the available data, whether explanations are validated rather than merely visualized, and whether published benchmarks support claims of transferability and deployment readiness. Across structured genotype-environment-management tables, satellite time series, UAV phenotyping, and multimodal pipelines, the literature does not support a universally superior architecture. Feedforward networks remain competitive when inputs are already distilled into informative agronomic descriptors; convolutional and recurrent models are strongest when spatial or temporal structure is explicit; and hybrid, attention-based, and multimodal systems are most persuasive when they solve a genuine alignment problem across modalities. The field is more mature in model innovation than in benchmark realism. Reported gains are frequently confounded by heterogeneous targets, leakage-prone splits, weak baselines, limited calibration analysis, sparse external validation, and inconsistent reporting across reviews. Explainability studies face a similar problem: SHAP, LIME, saliency maps, Grad-CAM, and attention mechanisms are useful only when fidelity, stability, and agronomic plausibility are evaluated together. For an AI audience, the central lesson is methodological rather than architectural: trustworthy deployment will depend less on deeper networks alone than on stronger benchmark ecosystems, uncertainty-aware reporting, realistic transfer tests, and clearer multimodal design.

Keywords: *crop yield prediction; deep learning; model-data fit; multimodal fusion; explainable AI; benchmarking*

Highlights

- *Reported deep-learning gains in crop-yield prediction are highly conditional on data regime, target definition, split design, and reporting quality.*

ARTICLE IN PRESS

- *Model choice should follow model-data fit: tabular, spatial, temporal, and multimodal settings reward different inductive biases.*
- *Explainability is useful only when fidelity, stability, and agronomic plausibility are evaluated together.*
- *Trustworthy benchmarking requires leakage-aware spatiotemporal splits, strong baselines, calibration analysis, and external validation.*
- *Future progress is most likely to come from benchmark-centered, multimodal, uncertainty-aware, and physics-guided research agendas.*

ARTICLE IN PRESS

ARTICLE IN PRESS

The abbreviations used throughout the manuscript are listed in Table 1.

Table 1: List of abbreviations

Abbreviation	Meaning	Abbreviation	Meaning
AI	artificial intelligence	Bi-LSTM	bidirectional long short-term memory
CNN-DNN	convolutional neural network–deep neural network	CNN-GRU	convolutional neural network–gated recurrent unit
CRPS	continuous ranked probability score	CY-Bench	benchmark dataset for subnational crop yield forecasting
EO	Earth observation	EVI	enhanced vegetation index
FPAR	fraction of photosynthetically active radiation	GAN	generative adversarial network
GNN	graph neural network	Grad-CAM	gradient-weighted class activation mapping
GRU	gated recurrent unit	LAI	leaf area index
LIME	local interpretable model-agnostic explanations	MAE	mean absolute error
ML	machine learning	MLP	multilayer perceptron
NLL	negative log-likelihood	PBM-RS-ML	process-based model–remote sensing–machine learning
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses	R^2	coefficient of determination
RGB	red, green, and blue	RMSE	root mean square error
SHAP	SHapley Additive exPlanations	SMOGN	Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise
UAV	unmanned aerial vehicle	VTCI	vegetation temperature condition index
BMA	Bayesian model averaging	XAI	explainable artificial intelligence

ARTICLE IN PRESS

1. Introduction

Crop yield prediction (CYP) is a core task in digital agriculture because reliable yield estimates inform food security planning, crop insurance, input allocation, logistics, market forecasting, breeding, and climate adaptation strategies. Yet producing forecasts that are not only accurate but also timely, robust, and actionable remains difficult [1], [2]. Yield emerges from interactions

ARTICLE IN PRESS

among biophysical and management processes, including weather variability, soil heterogeneity, genotype, agronomic practices, pests, and stress events, operating across multiple spatial and temporal scales. CYP is therefore not a simple regression task; it is a nonlinear, multimodal, and context-dependent inference problem [3], [4]. Conventional statistical and process-based crop models remain indispensable for agronomic analysis, but large-scale deployment can be limited by simplifying assumptions, labor-intensive calibration, and constraints on integrating heterogeneous data streams. At the same time, agriculture is generating large volumes of satellite imagery, UAV observations, weather reanalysis products, soil maps, phenological indicators, and farm records. These developments have brought deep learning to the forefront because it can learn hierarchical representations from minimally processed inputs and capture spatial, temporal, and cross-modal structure within a unified modeling pipeline [5], [6], [7].

Accordingly, the literature has expanded from early neural-network studies to a diverse ecosystem of multilayer perceptrons, convolutional neural networks, recurrent and temporal models, hybrid CNN-LSTM pipelines, attention mechanisms, transfer-learning strategies, and domain-adaptation methods [8], [9], [10]. Related advances in lightweight and feature-efficient CNNs from aerial detection, medical imaging, and biometric recognition also suggest useful design principles for resource-constrained UAV and edge-based agricultural applications, although their effectiveness must be validated directly in crop-yield settings [11],[12].

Recent studies also increasingly focus on explainability, reflecting the need to understand not only whether a model predicts accurately, but why it does so and whether that reasoning is agronomically plausible [13]. The central challenge is therefore no longer the absence of models, but the fragmentation of evidence across architecture choice, dataset regime, explainability practice, and evaluation design. Reported gains are often conditioned by crop, region, season, lead time, input modality, and validation strategy, making direct comparisons difficult. In many studies, architectural innovation is discussed without equal attention to dataset bias, domain shift, benchmark design, uncertainty reporting, or reproducibility. Consequently, strong performance on a single benchmark does not automatically translate into reliability, transferability, or decision usefulness in real agricultural settings. This review, therefore, offers an AI-centered critical synthesis of when architectural, explainability, and benchmarking claims in CYP are methodologically credible and when they are not. Its novel methodological contribution is a decision-oriented lens for selecting and evaluating deep-learning architectures according to the available data regime, expected deployment context, and validation design, rather than treating headline accuracy as sufficient evidence of model superiority.

2. Positioning Against Prior Reviews and Research Gap

Earlier reviews have been valuable in mapping the fast-growing landscape of crop-yield prediction, but most are stronger at cataloging methods than at judging the credibility of the evidence those methods produce. Broad machine-learning reviews provide historical context [3], [14]; remote-sensing-focused surveys clarify EO pipelines [4]; crop-yield-specific deep-learning reviews summarize architectures and application trends [15], [8], [13]; transformer and remote-sensing reviews track emerging attention-based models [9]; and agriculture-wide XAI reviews offer useful interpretability taxonomies [16]. What remains underdeveloped is the bridge among these strands: few reviews ask whether apparent model superiority still holds when data regime, explanation quality, and benchmark realism are examined together.Table 2 summarizes these differences.

ARTICLE IN PRESS

Table 2: Representative review articles and the specific gap addressed by this manuscript.

Review	Primary focus	What it covers well	What remains open
[3]	Broad ML crop-yield review	Algorithmic landscape, common input variables, historical baseline context	No deep-learning-specific explainability synthesis; limited attention to benchmark design and reproducibility
[4]	Deep learning and remote sensing for CYP	Remote-sensing pipelines, vegetation indices, satellites/UAVs, and seasonal imagery	Limited treatment of non-remote-sensing data, broader multimodal fusion, and XAI
[8]	Systematic review of deep learning for CYP	Architectures, crops, data sources, and application trends	Does not fully connect the dataset regime, explainability, and reproducibility
[13]	General deep-learning survey for CYP	Accessible overview of core model families, applications, and datasets	Limited critical comparison of validation protocols and benchmarking depth
[15]	Hybrid deep-learning models for CYP	Hybrid model families and application examples	Limited critical analysis of validation protocols, transferability, and evidence quality
[9]	Transformer and remote-sensing model review	Emerging attention-based architectures and broader remote-sensing context	Not yield-specific; limited scrutiny of agronomic transferability and benchmark realism
[16]	Agriculture-wide XAI review	Strong XAI taxonomy and cross-application perspective	Not yield-specific and does not link XAI to yield benchmarking or spatiotemporal splits
[17]	In-season yield forecasting review	Forecast timing, lead-time effects, input variables, and future research priorities	Not deep-learning-centered; limited integration of XAI, benchmark design, and model–data fit across regimes

Taken together, prior reviews establish the breadth of the field, but they leave unresolved a practical question that matters for both research and deployment: which deep-learning architectures are appropriate for specific crop-yield settings, given the available data, the required explanation burden, and the benchmark design used to support claims of reliability? This review addresses that question through an evidence-chain framework that treats model choice, dataset regime, target definition, explainability, validation design, and deployment readiness as interdependent. Unlike surveys that primarily catalogue architectures or applications, it evaluates whether the evidence supporting each reported improvement is credible for the claim being made: whether the architecture matches the data regime, the target supports comparison, the split supports the claimed transfer setting, explanations are validated rather than only visualized, and operational

ARTICLE IN PRESS

constraints are considered. Its contribution is therefore methodological and decision-oriented, helping readers select and evaluate models beyond headline accuracy. Accordingly, the review pursues five linked objectives:

i. To relate major deep-learning architecture families to the data regimes and agronomic questions they are best suited to address;
ii. To evaluate how datasets, sensing modalities, target definitions, and fusion strategies shape what can legitimately be compared across studies;
iii. To examine explainability methods in terms of what objects they explain, how faithful they are to model behavior, and how useful they are for agronomic reasoning;
iv. To identify recurring weaknesses in validation, uncertainty reporting, transferability, and reproducibility that undermine seemingly strong benchmark results;
v. To outline future directions that improve realism, transparency, and transportability simultaneously, rather than optimizing only headline accuracy.

3. Review Approach and Scope

This manuscript is a structured critical review rather than a registered systematic review or meta-analysis [18]. It was guided by PRISMA 2020 reporting principles [19] to improve transparency in searching, screening, eligibility assessment, and reporting; however, its purpose is critical synthesis rather than pooled quantitative effect estimation. The review focuses on model-data fit, validation design, explainability, benchmarking, and deployment relevance.

3.1 Review Question, Pilot Search, and Search Strategy

The review was guided by the following question: Which machine-learning and deep-learning models, input data sources, predictor variables, and evaluation practices have been used for maize or crop-yield prediction, and what methodological gaps remain for trustworthy, transferable, and deployment-ready crop-yield forecasting?

A pilot search was first conducted to refine the scope of the review and identify the most relevant search terms. The search covered the period 2017–2026, because recent studies were expected to contain the most mature applications of machine learning and deep learning for crop-yield forecasting. The main databases searched were Google Scholar, ScienceDirect, and Wiley Online Library. Backward snowballing was also conducted by screening the reference lists of eligible papers and relevant review articles. Where database interfaces allowed field restrictions, searches were conducted across titles, abstracts, and keywords.

The final Boolean search string was:

("machine learning" OR "deep learning" OR "artificial intelligence" OR XGBoost OR "random forest" OR "support vector regression" OR "neural network") AND ("yield prediction" OR "yield forecasting" OR "yield estimation") AND (maize OR corn OR "crop yield") AND ("remote sensing" OR satellite OR "vegetation index" OR NDVI OR rainfall OR temperature OR "soil moisture").

Search results were exported into a spreadsheet and each record was assigned a unique identifier. Duplicate records were removed before title-and-abstract screening. Full texts of potentially eligible studies were then assessed against predefined eligibility and quality criteria.

ARTICLE IN PRESS

3.2 **Selection, Eligibility, and Quality Assessment of Primary Studies**

A total of 1,859 records were retrieved during the identification stage. After removal of duplicates and clearly irrelevant records, 908 records were retained for title-and-abstract screening. Application of the inclusion and exclusion criteria reduced the set to 489 full-text articles assessed for eligibility. Following the quality assessment, 100 studies were retained for final data extraction and synthesis. The selection process is summarized in Figure 1 using a PRISMA-style flow-diagram structure, which reports the number of records identified, screened, assessed for eligibility, and included in the final synthesis.

The inclusion criteria were as follows: peer-reviewed full-text journal or conference article; publication in English between 2017 and 2026; application of machine learning, deep learning, or comparable statistical learning to maize, corn, or crop-yield prediction, estimation, or forecasting; clear reporting of data source, predictor variables, model type, and at least one performance metric; and relevance to remote-sensing, climatic, soil, management, genotype, or multimodal drivers of yield. The exclusion criteria were as follows: non-English papers; editorials, theses, presentations, preprints, or non-peer-reviewed reports; studies without yield as a prediction target; studies not using machine learning, deep learning, or comparable predictive modeling; papers without sufficient methodological detail; duplicate studies; and studies for which the full text was not accessible.

Study quality was assessed using a structured appraisal checklist. Each study was examined for clarity of prediction target, adequacy of data-source description, reporting of predictor variables, appropriateness of model family, clarity of validation design, leakage-control measures, use of baseline models, reporting of performance metrics, treatment of uncertainty or explainability where applicable, and reproducibility information such as preprocessing steps, code, data availability, or split logic. Studies with clearer target definitions, stronger validation protocols, leakage-aware split designs, external testing, ablation analysis, or reproducibility artifacts were given greater interpretive weight in the synthesis.

ARTICLE IN PRESS

ARTICLE IN PRESS

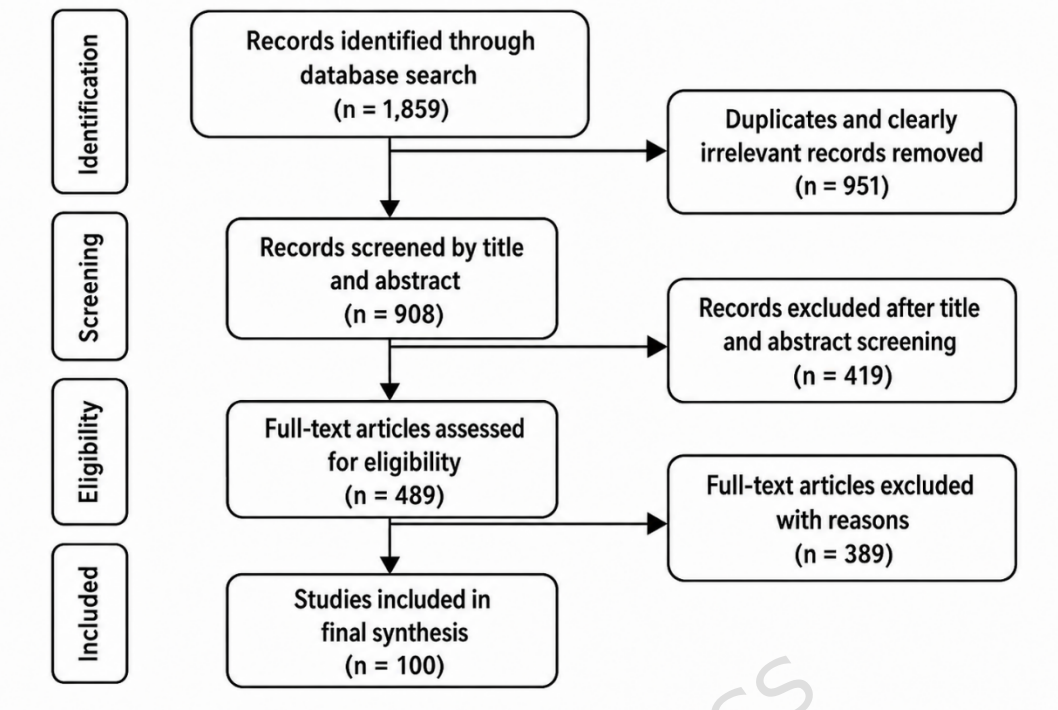


Figure 1. PRISMA-style flow diagram for study identification, screening, eligibility assessment, and final inclusion.

3.3 Analytical Framework for Critical Synthesis

To avoid treating models, datasets, explainability methods, and benchmarks as separate topics, this review applies an evidence-chain framework for evaluating trustworthy crop-yield prediction. The framework is based on five linked questions. First, what data regime does the study address i.e. structured agronomic tables, satellite time series, UAV or proximal imagery, or multimodal fusion? Second, does the selected architecture match the structure of the data and the prediction target? Third, does the validation design support the generalization claim being made, particularly with respect to spatial, temporal, cross-year, or external transfer? Fourth, when explainability methods are used, are the explanations validated for fidelity, stability, and agronomic plausibility rather than presented only as visual outputs? Fifth, does the study provide evidence of deployment readiness, including uncertainty reporting, missing-data handling, data latency, reproducibility, and integration with operational decision workflows?

This framework is used consistently across the review. Section 4 evaluates how data regimes shape what models can legitimately learn. Section 5 assesses whether architecture families are appropriate for those regimes. Section 6 examines whether explainability methods provide validated evidence or only descriptive interpretation. Section 7 evaluates whether benchmark and split designs support the claims made about transferability and reliability. Section 8 then considers whether the resulting models are sufficiently robust, transparent, and operationally feasible for digital agriculture deployment.

4. Dataset Regimes, Multimodal Fusion, and Data Augmentation

ARTICLE IN PRESS

Deep learning in crop-yield prediction is often discussed as though architecture alone determines performance. That framing is analytically weak. In practice, what a model can learn, how early it can predict, and whether its error estimates can be trusted depend first on the dataset regime: the spatial unit, forecast horizon, label construction, and combination of modalities available for training. County-scale satellite forecasting, breeding-program tabular prediction, and plot-scale UAV mapping are therefore not variants of the same task but distinct prediction problems with different signal-to-noise ratios, transfer risks, and interpretation burdens [20], [21]. Table 3 summarizes the main regimes and their recurring methodological risks.

Table 3: Main crop-yield data regimes and their recurring methodological risks.

Refs.	Data regime	Typical inputs	Scale	Main strengths	Main risks
[5], [22]	Structured genotype-environment-management tables	Pedigree or hybrid identifiers, weather summaries, soil descriptors, and management records	Plot, trial, breeding program	Direct agronomic covariates and strong relevance to breeding decisions	Proprietary access, environmental imbalance, weak spatial context, and inconsistent reporting
[23], [24]	Satellite time series and vegetation indices	MODIS, Landsat, Sentinel, vegetation indices, climate time series	County, regional, national	Scalable coverage, earlier in-season prediction, and policy relevance	Cloud gaps, spatial autocorrelation, mixed pixels, and label mismatch with administrative yields
[20], [21]	UAV, multispectral, hyperspectral, and phenotyping imagery	RGB, multispectral, hyperspectral, thermal, canopy traits	Plot, field	High spatial detail and sensitivity to canopy condition	Small samples, expensive labels, and limited cross-site generalization
[25], [26], [27]	Multimodal fusion	EO plus weather, soil, management, or genotype data	Field to regional	Closest match to real agronomic complexity and multi-scale prediction	Alignment and missingness problems, ad hoc fusion choices, and heavy pre-processing dependence

4.1 Data Regimes and Prediction Settings

4.1.1 Structured genotype-environment-management tables

In structured genotype-environment-management settings, such as breeding programs or performance trials, inputs often include genotype identifiers or pedigree signals, weather

ARTICLE IN PRESS

summaries, soil properties, and management descriptors, with yields measured under relatively controlled protocols [5], [22]. These datasets are well-suited for feedforward or sequence models because the main challenge is learning nonlinear interactions among agronomic variables. But their apparent cleanliness can also mask limited environmental coverage and poor transportability: models may learn site- or program-specific regularities that do not generalize outside the breeding network that generated the data. A strong result in this regime is therefore not evidence that the same architecture will perform equally well in open operational forecasting.

4.1.2 Satellite Time Series and Regional Forecasting

Satellite-based yield prediction has driven much of the field's progress toward operational scale because repeated observations over large geographies enable pre-harvest estimation. Foundational county-level studies on soybean and corn/soybean forecasting showed how CNN-, LSTM-, and transfer-learning-based architectures can exploit remotely sensed seasonal trajectories at scale [23], [24]. Later studies that integrate climate, soil, and satellite data for wheat or rice across China show why this regime remains attractive to governments and market analysts: it supports broad coverage, earlier forecasts, and lower per-unit data cost. However, the regime is often over-interpreted. Administrative yield labels smooth over cultivar choice, planting dates, stand density, and within-field stress, so a county-level model may achieve high accuracy while learning regional averages rather than agronomically actionable field processes. Spatial autocorrelation further inflates apparent generalization when neighboring counties or repeated environmental conditions are split carelessly between training and test sets. The implication is important for this review: success at the county scale should not be presented as evidence that the same architecture is suitable for precision agriculture. Sub-field studies show that performance becomes substantially harder when the task shifts from county-level interpolation to field-level prediction under stronger spatial controls [28], [29].

4.1.3 UAV, Multispectral, Hyperspectral, and Proximal Phenotyping Imagery

UAV, multispectral, hyperspectral, thermal, and proximal sensing regimes move CYP closer to plant physiology and management action. Studies on multitemporal UAV imagery, multimodal soybean prediction, spatially rigorous corn prediction, and rice prediction under spatial heterogeneity show that high-resolution sensing can capture canopy structure, temperature, and fine spatial variability that coarse satellite products cannot [20], [21]. This makes UAV regimes especially relevant for breeding plots, irrigation experiments, and within-field decision support. Yet proximity to plant biology does not eliminate statistical fragility. Many UAV studies still rely on one or two seasons, dozens to hundreds of plots, and tightly controlled experiments. That design is useful for mechanism discovery, but risky for broad claims of model generality. Killeen et al. [30] are especially instructive because they showed that spatial cross-validation yields more realistic performance than conventional random folds in corn yield mapping. Likewise, Zhou et al. [31] showed that accounting for spatial heterogeneity can materially change both performance estimates and the apparent ranking of models. A critical reading must therefore distinguish physiological sensitivity from operational robustness: a model that detects fine-grained canopy patterns is not automatically a model that will transfer across seasons, sites, or germplasm.

4.1.4 Multimodal Data Fusion

Multimodal data fusion has become central because yield is jointly shaped by canopy status, weather exposure, soil constraints, management decisions, and, in some settings, genotype. From

ARTICLE IN PRESS

an agronomic perspective, fusion is attractive because no single modality observes both stress exposure and crop response. Weather indicates the conditions a crop experienced; spectral and thermal data show how the canopy actually responded; soils and management constrain both [20], [27]. When fusion works, it matches the logic of the biological system better than any single-source model. However, multimodality is not automatically beneficial. Every added modality introduces new assumptions about temporal alignment, spatial support, missingness, pre-processing, and measurement error. A fusion model may look sophisticated while simply concatenating poorly matched features. The key question is not whether more data are available, but whether the additional modality contributes non-redundant information at the same decision scale as the label.

Several crop-yield studies demonstrate the value of complementary modalities. Maimaitijiang et al. [20] showed that fusing RGB, multispectral, and thermal UAV data improved soybean yield prediction relative to simpler alternatives. Mia et al. [32] combined UAV multispectral imagery with weather data for rice and similarly found that multimodal models improved predictive accuracy, while also noting that models with similar summary accuracy could still produce different spatial yield patterns. More recent adaptive-fusion work at the sub-field scale reaches a similar conclusion: multimodal gains are strongest when the fusion strategy respects the spatial and temporal complementarity of the inputs [33].

Large-scale studies point in the same direction, but they also expose the main unresolved problem. Global and subnational systems that combine satellite, climate, soil, and related environmental predictors are most persuasive when the modalities capture distinct agronomic mechanisms [28], [34]. Yet temporal mismatches between imagery and weather windows, unequal spatial supports across soil and satellite products, and non-random missingness caused by cloud cover or sensor failure can materially change the results. Accordingly, future crop-yield benchmarks should report modality-alignment rules, missing-data handling, forecast horizon, ablation results, spatial and temporal holdout design, and uncertainty behavior alongside raw accuracy [33], [29].

ARTICLE IN PRESS

4.2 Cross-Regime Implications

Across regimes, meaningful model comparison begins with an explicit prediction setting. A county-scale end-of-season forecast from public satellite archives, a within-program breeding prediction from structured G × E × M tables, and a plot-scale UAV map are not directly comparable simply because all report RMSE or R^2. The regime determines target difficulty, the likely sources of leakage, and the extent to which results are actionable for breeding, insurance, or farm management. It also determines the burden of proof: structured tables invite feature-attribution and interaction analysis; satellite models must show that predictions are not driven by clouds, mixed pixels, or location proxies; and UAV models must demonstrate robustness to stage, lighting, and plot layout. These regime differences lead directly to two under-examined issues in the CYP literature: how yield labels are constructed and how limited-data problems are handled.

4.3 Yield Labels and Data Augmentation

The quality of the yield label is just as important as the quality of the model. The literature can appear more definitive than it really is because many papers report metrics without clearly defining what 'yield' means operationally. Depending on the study, the target may be plot-harvest grain mass per hectare, bushels per acre after moisture correction, production aggregated over administrative units, or a normalized yield index. These targets are not interchangeable, and neither

ARTICLE IN PRESS

are the errors reported against them [35], [36]. A model trained on consistent plot-harvest measurements, therefore, solves a fundamentally different problem from one trained on county statistics that average over heterogeneous fields and reporting practices.

This label issue becomes especially important when authors compare results across spatial scales or forecast horizons. Aggregated county labels can support early warning and policy planning, but they blur field-level variability and can reward models that track large-scale climate or planting-area effects rather than physiological crop status [7], [25]. Plot-level labels provide tighter biological coupling to the sensed canopy but are often season-, site-, or genotype-specific [20]. Accordingly, survey-level interpretation should always consider accuracy metrics alongside label provenance, aggregation level, harvest protocol, and lead time; otherwise, apparently precise comparisons can be methodologically misleading.

These constraints help explain why augmentation is attractive in crop-yield studies. Image-based pipelines use cropping, rotation, flipping, radiometric perturbations, and biophysical-index augmentation to enlarge the effective training set [37], while GAN-based augmentation and imbalance-oriented resampling are used to counter small-sample and long-tail problems [38], [39]. But augmentation is not automatically benign. Some transformations preserve agronomic invariants; others distort canopy geometry, phenology, or sensor-response patterns in ways that make the model better at the synthetic task than at the real one. The apparent gain can therefore come from regularization against a narrow benchmark rather than from improved agronomic generalization.

Recent wheat and maize studies highlight both the potential and the limits of augmentation. Data augmentation using remotely sensed biophysical indices has improved deep neural network estimates under limited-sample conditions [37]. Related work also suggests that GAN-based augmentation [38] and tabular resampling strategies such as SMOGN can help in data-scarce or imbalanced settings [39], [40], [41]. However, augmentation does not introduce new agronomic mechanisms or repair mismatches between observed labels and predictor variables. These gains are convincing only when evaluation protocols enforce strict spatial and temporal separation between training and test data and when generated samples remain agronomically plausible.

5. Deep-Learning Architectures and Model-Data Fit

Deep-learning architectures differ more in their inductive bias than in prestige[42]. In crop-yield prediction, model–data fit refers to how well an architecture's assumptions about spatial, temporal, or cross-modal structure match the organization of the inputs and the scale of the target. CNNs are natural when local spatial patterns matter; recurrent models are natural for ordered seasonal sequences; transformers are appealing when long-range interactions and flexible cross-feature weighting matter; and feedforward models remain competitive when agronomic descriptors already summarize the season effectively [43], [44]. When architecture and data are well matched, learning is more efficient, and generalization is usually stronger. When the fit is poor, models can overfit, train inefficiently, or learn shortcuts that do not survive deployment.

5.1 Feedforward Networks for Structured Agronomic Data

Feedforward networks and multilayer perceptrons remain stronger baselines in crop-yield prediction than the current fashion for novel architectures sometimes implies. In large structured agronomic datasets, especially those dominated by genotype, weather, soil, and management

ARTICLE IN PRESS

descriptors, their flexibility is often sufficient to capture non-additive interactions and threshold effects [5], [22]. This is an important corrective to the common assumption that simple tabular models are necessarily obsolete once deep learning enters the conversation. If the inputs already summarize the season meaningfully, a well-regularized MLP may match or exceed more elaborate spatiotemporal models.

Their limitation is not that they cannot learn, but that they must learn without much built-in structure. They possess little explicit temporal or spatial inductive bias, so their success often depends on the feature engineering done upstream. Aggregated weather variables, one-hot genotype identifiers, or site encodings can simplify the prediction problem, but they can also encourage shortcut learning and hidden leakage if the same genotypes, trials, or environments recur across splits [5], [45]. A strong result from a feedforward model should therefore be interpreted as evidence that the dataset has been meaningfully summarized, not as proof that time and space are irrelevant.

5.2 Convolutional Networks for Imagery and Gridded Inputs

Convolutional neural networks are natural choices when the input is image-like or spatially structured. Their local receptive fields and weight sharing make them effective at extracting canopy texture, spectral contrasts, and neighborhood patterns from UAV imagery, raster stacks, and hyperspectral measurements [20]. Recent lightweight CNN developments offer useful design lessons for image-based agricultural applications. Direction-aware and anchor-free aerial detection models show how compact feature extraction and efficient fusion can preserve spatial detail while reducing computational cost, while lightweight biometric CNNs and KACNet demonstrate efficient feature representation and real-time inference. Although developed outside crop-yield prediction, these ideas are relevant to UAV phenotyping, field imagery, and edge-based agricultural systems, but they still require validation on agricultural datasets using strong baselines and realistic spatiotemporal splits [44], [45]. In crop-yield prediction, CNNs are most convincing when the spatial structure in the input genuinely corresponds to the spatial scale at which yield-relevant signals are expressed. This is often true at the plot level, where canopy architecture, lodging, or stress signatures can be visible in the imagery itself. The problem is that many agricultural targets are not defined at the same scale as the image features being learned. County-level yield labels, for example, summarize large heterogeneous areas, so a CNN trained on gridded or tiled imagery may be learning local proxies for a spatially aggregated target [7], [23]. Mixed pixels, cloud contamination, varying revisit frequency, and cross-sensor harmonization further complicate interpretation. Accordingly, CNN success should not be read as evidence that spatial texture alone explains yield; it may simply be the most efficient way to compress correlated remote-sensing signals under a particular benchmark. When temporal evolution or multimodal dependency is central, CNNs often need to be combined with other mechanisms rather than used alone.

5.3 Recurrent Models for Seasonal Dynamics

Recurrent models such as LSTMs and GRUs have become prominent because crop growth is inherently sequential. As weather stress accumulates, vegetation trajectories evolve, and the same variable can have different effects at different growth stages. Studies using CNN-RNN frameworks, soybean county-level CNN-LSTM models, recurrent 3D convolutional pipelines over multitemporal imagery, and optimized LSTM variants demonstrate why temporal encoders have become attractive for in-season forecasting [24], [46], [47], [48]. Their core advantage is not mere

ARTICLE IN PRESS

sequence handling, but stage sensitivity: they can represent the fact that a heat shock during flowering is not equivalent to the same temperature anomaly early in vegetative growth.

Even so, recurrent performance is not automatically a sign that the model has learned agronomic dynamics well. Sequence models are sensitive to how time steps are defined, how missing observations are imputed, and whether season boundaries are aligned across regions and crops. A recurrent model can also memorize calendar regularities that appear predictive in familiar years but degrade under climate anomalies or under different crop calendars. For this reason, strong evaluation of LSTM or GRU systems should include ablations on sequence length, robustness to missing time steps, and external testing across years or environments [24], [46], [2].

5.4 Hybrid Spatiotemporal Architectures

Hybrid spatiotemporal architectures, such as CNN-LSTM, ConvLSTM, CNN-GRU, or ensemble combinations of convolutional and tabular branches, currently represent the most mature empirical compromise for many real-world yield problems [24], [49]. They are attractive because they align with the task structure: spatial encoders compress imagery or gridded inputs, temporal modules track seasonal evolution, and auxiliary dense branches absorb weather, soil, or management covariates. When yield depends on both where and when stress develops, hybrids embody a plausible inductive bias rather than complexity for its own sake.

That said, hybrid models are often reported as though the architecture itself were the source of their superiority, when in practice their advantage may come from more careful feature alignment or simply from ingesting more modalities. Ensemble convolutional neural network-deep neural network (CNN-DNN) systems for corn yield [45], [46] and transfer-learning frameworks that combine remote-sensing representations across crops [21] are informative precisely because they show that performance depends on the interaction between architecture and input design. A rigorous interpretation, therefore, requires ablations that separate spatial encoding, temporal encoding, transfer learning, and multimodal fusion effects. Without those controls, the literature risks rewarding engineering bundle size rather than true model insight.

5.5 Attention and Transformer-Inspired Models

Attention mechanisms and transformer-inspired encoders extend the search for better inductive bias to settings where long-range temporal dependence and adaptive variable reweighting matter. In crop-yield and breeding applications, attention layers can reweight influential time windows, sensors, or feature groups instead of compressing all observations through a uniform recurrent bottleneck [50], [51]. Transformer-style architectures further suggest that long temporal contexts and cross-variable interactions can be modeled without recurrence, which makes them appealing for multimodal and multivariate prediction tasks [52], [53].

Their appeal is therefore genuine: they offer scalable sequence modeling and a flexible way to represent heterogeneous dependence. Even so, claims of superiority should be interpreted cautiously. Reported gains often come from relatively narrow or task-specific datasets, and comparisons are not always supported by external validation, matched hyperparameter budgets, or sufficiently broad benchmarking. Attention weights should also not be treated as explanations by default; they may indicate which inputs the model emphasizes, but they do not uniquely establish causal influence or stable feature importance [54], [55].

5.6 Graph Neural Networks

ARTICLE IN PRESS

Where relational dependence is central, graph neural networks (GNNs) are especially attractive because yield prediction may depend on relations among fields, counties, or genotypes rather than on isolated samples. Instead of treating each unit independently, GNNs represent units as nodes and propagate information across edges defined by adjacency, similarity, or other agronomically motivated links. This makes them potentially useful for spatially connected administrative units, breeding networks, and settings with missing or partially observed attributes. Recent studies show early promise in genotype-topological and trait-missing settings [56], [57], [58], while related temporal-geospatial frameworks highlight the value of explicitly representing spatial dependence [59]. However, GNNs also introduce distinct sources of fragility. Performance depends heavily on graph construction, neighborhood design, and whether the chosen edges reflect agronomic processes rather than mere spatial convenience. They can also suffer from over-smoothing or unstable message passing as graphs deepen [60]. Accordingly, GNNs are promising for explicitly relational crop-yield settings, but their evidence base remains smaller and less standardized than that for CNN-, LSTM-, or hybrid models [59], [56], [58].

A related direction is graphical or correlation-structure learning for representing dependencies among yield drivers. Crop-yield datasets often contain strongly correlated predictors, including rainfall, temperature, evapotranspiration, vegetation indices, soil moisture, phenological variables, and management proxies. Graphical Gaussian models and related structure-learning methods can estimate conditional dependencies and help distinguish direct associations from redundant or proxy signals [61]. In crop-yield prediction, these structures could support feature grouping, multimodal fusion, causal hypothesis generation, and graph construction for GNNs. They may also reveal whether a model relies on stable agronomic relationships or on correlations that are likely to fail under spatial, temporal, or climatic shift. Table 4 summarizes the main architecture families and their best-fit data regimes.

Table 4: Deep-learning architecture families, the data regimes they fit best, their main strengths, and recurring limitations in crop-yield prediction.

Refs.	Model family	Best-fit data regime	Main strengths	Main limitations
[5], [22]	DNN or MLP	Large structured tabular datasets with genotype, weather, soil, and management summaries	Flexible for heterogeneous features and strong on breeding-style records	Weak explicit temporal or spatial inductive bias; relies on aggregated features
[20], [21]	CNN	Imagery, raster stacks, UAV, and spatially structured remote-sensing inputs	Efficient local spatial feature extraction	Limited long-range temporal modeling when used alone
[23], [47]	LSTM or GRU	Seasonal weather, vegetation index, or management sequences	Captures temporal dependence and phenological timing	Training complexity, sensitivity to sequence design, and weaker parallelization

ARTICLE IN PRESS

[24], [49]	Hybrid CNN-LSTM, CNN-GRU, or ensemble systems	Multimodal spatiotemporal prediction problems	Combines spatial and temporal inductive bias; empirically mature in practice	Fusion strategies are often ad hoc and not always well ablated
[27], [53]	Attention or transformer-inspired models	Long sequences, feature reweighting, and multimodal interaction settings	Captures long-range dependencies and flexible cross-feature interactions	The evidence base is still narrow; often single-region and under-validated externally
[56]	Graph neural networks	Relational, spatial-topological, or attribute-missing data	Can exploit inter-field or inter-county relations	Sensitive to graph design and less standardized in benchmarking
[55], [57], [58], [59]	Bayesian, self-supervised/transfer, and physics-guided models	Risk-sensitive, label-scarce, and extrapolation-prone applications	Add uncertainty, label efficiency, or agronomic structure	Underrepresented, harder to benchmark fairly, and more complex to implement

To complement the architecture-level comparison in Table 4, Table 5 summarizes representative studies by dataset or prediction setting, model family, metrics, and major contribution. The table is not intended as a direct ranking because the studies differ in crop, spatial scale, target definition, forecast horizon, validation protocol, and evaluation metric; rather, it shows how reported findings depend on the interaction between data regime, architecture, and evaluation design.

Table 5: Comparative summary of representative crop-yield prediction studies.

ARTICLE IN PRESS

ARTICLE IN PRESS

Reference	Region	Data regime / input features	DL/ML models	Reported best performance or major finding
[60]	China	Remote sensing, drought, climate, soil moisture, SIF, GPP, FPAR, LAI	RF, XGBoost, CNN, CNN-LSTM-Attention	CNN-LSTM-Attention achieved strong maize-yield performance: R^2 = 0.80; MAPE = 4.21%
[38]	India	Plant traits, canopy temperature, chlorophyll readings	Stacked CNN ensemble with DT, XGBoost, and LASSO	Stacked CNN ensemble achieved R^2 = 0.9165; MSE = 0.6893
[61]	Senegal	Climate, drought indices, vegetation index, growing-degree indicators	XGBoost, RF, CatBoost, LightGBM	CatBoost performed best: RMSE = 0.330; NRMSE = 0.093; MAE = 0.17
[24]	USA	MODIS surface reflectance and land-surface temperature	Ridge, LASSO, RT, RF, DFNN, 3D-CNN, YieldNet	YieldNet outperformed baselines: maize RMSE = 17.49; soybean RMSE = 5.07
[62]	China	MODIS reflectance, climate forcing, soil data, DEM	RF, XGBoost, LSTM, LSTM-DGP, hybrid CNN-MLP-GRU	Hybrid model outperformed baselines: RMSE = 1006 kg/ha; MAPE = 17.1%
[46]	USA	MODIS, Daymet weather, and soil variables	CNN, LSTM, CNN-LSTM, LASSO, RF, Ridge, DNN	In-season models achieved R^2 around 0.73–0.75 and MAPE around 7.93–8.11%
[33]	USA	Vegetation indices, precipitation, temperature, LST, shortwave radiation	RF, DNN, ADANN, PDANN	PDANN performed best in 2020: R^2 = 0.71; RMSE = 0.83 t/ha
[63]	USA	Soil, weather, moisture, temperature,	GNN-RNN, LSTM, GRU, CNN, CNN-	GNN-RNN outperformed alternatives: RMSE

ARTICLE IN PRESS

		vapor-pressure variables	RNN, GNN, regressors	= 0.5836; R^2 = 0.6591; PCC = 0.8259
[64]	USA	Meteorological spatial-temporal variables	MMST-ViT, ConvLSTM, CNN-RNN, GNN-RNN	MMST-ViT achieved corn RMSE = 10.5; R^2 = 0.811; Corr = 0.900
[65]	South Africa	FPAR, soil moisture, precipitation, temperature, solar radiation	TabPFN, SVR, GPR, GBR, RF, LASSO	ML approach and TabPFN variants achieved rRMSEp values of 7.39–9.37%
[66]	China	Vegetation indices, SIF, LST, climate, soil texture, SOC, socioeconomic variables	CNNAtBi-GRU, CNN-GRU, GRU, LightGBM, LASSO	CNNAtBi-GRU performed best: R^2 = 0.896; RMSE = 908 kg/ha
[67]	Egypt	Vegetation indices, climate, solar radiation, soil properties	H2O AutoML, GBM, XGBoost, RF, DNN, GLM, stacked ensembles	H2O AutoML stacked ensemble achieved R^2 = 0.51, d = 0.82, RMSE = 1.22, MAE = 0.85
[68]	Kenya, Somalia, Burkina Faso, Malawi	Precipitation, NDVI, temperature, evapotranspiration	Extremely Randomized Trees	Achieved NSE > 0.6 and MAPE < 20% months before harvest
[69]	Tanzania	NDVI, temperature, rainfall, soil moisture	Unimodal and bimodal LSTM	MAPE = 3.656% for unimodal LSTM and 6.648% for bimodal LSTM
[70]	Ethiopia, Kenya, Malawi, Tanzania, Nigeria	NDVI, EVI, LST	LSTM, Ridge, LSTM-GP	LSTM-GP achieved R^2 = 0.5 and r = 0.75 on combined country test
[71]	Uganda	Vegetation indices, fertilizer use, seed adoption, soil and plot variables	MLR, linear regression, Scalable Crop Yield Mapper	Satellite-based estimates achieved R^2 = 0.54

ARTICLE IN PRESS

[72]	Uganda	Weather, CO_2, soil, planting/harvest dates, density, fertilizer	DSSAT-CERES	Achieved $R^2 = 0.78$ and $d = 0.63$
[73]	Uganda	Weather, ET_0, CO_2, soil, planting dates, maturity, harvest index, density	AquaCrop, linear regression	AquaCrop achieved $R^2 = 0.92$, NSE = 0.91, RMSE ≈ 0.09 t/ha

This comparative summary shows why reported performance should be interpreted cautiously. High R^2 or low RMSE values are not directly comparable across studies because the reviewed papers differ in region, crop, spatial unit, input features, target definition, and validation strategy. The table also shows that strong results occur across several model families, including tree-based ensembles, CNN and LSTM hybrids, transformers, graph-based models, domain-adaptation models, and process-based crop models. This reinforces the central argument of the review: model suitability depends on data regime, validation design, and deployment context rather than architecture novelty alone.

5.7 Uncertainty-Aware, Self-Supervised, and Physics-Guided Directions

Recent reviews show that crop-yield prediction has shifted from classical regression and tree-based models toward multimodal deep learning that combines satellite imagery, weather, soil, and management variables [3], [4]. Supervised baselines such as deep neural networks, CNN-RNN hybrids, and temporal models have improved point accuracy [5], [6], but harder evaluations reveal clear degradation under out-of-year, out-of-region, and field-to-field transfer [25], [33]. From an AI perspective, the key question is therefore not only whether these models reduce average error, but whether they improve reliability, label efficiency, and robustness under distribution shift. This section examines uncertainty-aware, self-supervised, transfer-based, and physics-guided approaches because they directly address three persistent weaknesses in accuracy-driven studies: overconfidence, dependence on expensive labels, and weak extrapolation under spatial, temporal, and climatic shift.

ARTICLE IN PRESS

5.7.1 Uncertainty-Aware Crop Yield Prediction

Uncertainty-aware modeling is especially relevant in agriculture because downstream decisions are asymmetric and risk-sensitive: farmers, insurers, traders, and food security planners need not only expected yields but also the probability of adverse outcomes. Early work by You et al. [7] demonstrated how deep Gaussian processes can combine learned representations with predictive distributions, and later work extended this agenda by applying Bayesian neural networks to corn yield uncertainty analysis [55]. The key distinction is between aleatoric uncertainty, which captures irreducible environmental variability, and epistemic uncertainty, which captures model ignorance caused by limited, noisy, or shifted data. In deployment settings, epistemic uncertainty is often the more consequential signal because it indicates when the model is operating beyond the support of the training distribution.

These methods are not interchangeable. Gaussian processes and Bayesian deep models are attractive when nonlinear spatiotemporal learning and predictive distributions are both required [7], [55]. Their main contribution is not simply to add another output statistic, but to shift crop-yield forecasting toward decision-relevant reliability. However, the literature still suffers from

ARTICLE IN PRESS

calibration problems. Reporting intervals or predictive variances is not sufficient unless those quantities are evaluated with empirical coverage, reliability diagrams, negative log-likelihood, or other proper scoring rules [35], [36]. Many studies also assess uncertainty on regions or years that closely resemble the training data, so apparent confidence can mask brittleness under a genuine distribution shift. The stronger conclusion is not that probabilistic models are already deployment-ready, but that they provide the right framework for asking when predictions should and should not be trusted. Future work should report point accuracy and calibration together, separate aleatoric from epistemic behavior when feasible, and test uncertainty in extreme-weather years, data-poor regions, and other out-of-distribution settings.

Uncertainty-aware prediction should also include model uncertainty, not only parameter or predictive uncertainty within a single architecture. Bayesian Model Averaging (BMA) and related ensemble-based Bayesian approaches provide a principled way to combine competing model classes by weighting their predictions according to posterior model support or predictive evidence [74]. This is especially relevant to crop-yield prediction because the literature does not support a universally superior architecture across structured tables, satellite time series, UAV imagery, and multimodal settings. Instead of selecting a single "best" model, BMA can combine, for example, tabular models, CNN-based image encoders, recurrent or transformer-based temporal models, and physics-guided predictors while propagating uncertainty about which model family is most reliable under a given data regime. In deployment settings, this can be useful when different architectures perform better in different regions, seasons, crops, or input-modality conditions. However, BMA is not a substitute for rigorous validation as model weights must be estimated on leakage-free calibration data, and ensemble uncertainty should be evaluated with proper scoring rules, calibration diagnostics, and external spatiotemporal tests.

5.7.2 Self-Supervised and Transfer-Learning Strategies for Label-Efficient Prediction

Label scarcity remains a structural bottleneck in crop-yield prediction, especially when fine-scale yield records are sparse, delayed, or inconsistently collected. Transfer learning addresses this constraint by adapting representations learned in data-rich settings. Khaki et al. [24] showed that deep transfer learning can support simultaneous corn and soybean prediction from remote-sensing data, while Ma et al. [33] used partial domain adaptation to improve transferability under distribution shift. Self-supervised learning is related but distinct. Instead of relying on source-task labels, it learns general spectral, spatial, or temporal structure from unlabeled imagery, climate series, or other sensor data before fine-tuning on scarce yield observations [75]. This distinction matters because transfer gains depend heavily on source-target similarity, whereas self-supervised gains are valuable precisely when labeled target data are limited.

Transfer learning currently has a broader empirical base than crop-specific self-supervised learning [24]. Early crop-oriented self-supervised results are encouraging, but the evidence base remains much smaller [57], [75]. Claims of label efficiency are strongest when models are tested across regions, years, scales, and label budgets under matched lead times and spatially disjoint splits. Otherwise, apparent gains can simply reflect shared geography, crop calendars, or sensor artifacts between training and test data [25], [76].

These approaches are strategically important for regions with rich remote-sensing archives but weak yield reporting. Yet negative transfer remains a central risk: features learned from one crop calendar, management system, or sensing regime may not hold elsewhere. The literature, therefore, supports a cautious conclusion. Transfer-based and self-supervised methods are promising for

ARTICLE IN PRESS

data-poor environments, but their value must be demonstrated under realistic out-of-region, out-of-year, low-shot, and missing-modality settings rather than inferred from random cross-validation alone [25], [77].

5.7.3 Physics-Guided and Hybrid Process-ML Modeling

Physics-guided and hybrid models address a key weakness of purely data-driven systems: a strong historical fit may not survive novel weather, management, or stress conditions. Agronomic knowledge can enter through crop-model states, process-informed features or priors, structured latent states, data assimilation, or mechanistic loss constraints. Because these mechanisms are not equivalent, studies should state precisely how guidance is introduced. Evidence from the U. S. Corn Belt and recent physics-guided deep learning indicates that process information can improve plausibility and robustness [58], [59].These benefits are not guaranteed. Mis-specified priors, costly crop-model calibration, and errors inherited from either component can offset gains, while richer inputs or heavier expert tuning may confound comparisons. Physics-guided models should therefore be evaluated against matched data-driven baselines, with transparent ablations showing whether improvements arise from the process model, additional features, architecture, or loss design. Table 6 compares these approaches and their evaluation priorities [58], [59].

Table 6: Comparative assessment of uncertainty-aware, self-supervised/transfer, and physics-guided approaches to crop-yield prediction.

Approach	Primary value	Current evidence	Main limitation	Evaluation priority
Uncertainty-aware	Quantifies confidence and flags extrapolation risk.	Established methodological base: deep Gaussian processes and Bayesian neural networks, [55].	Often under-calibrated; uncertainty is rarely tested under a genuine distribution shift.	Report coverage, reliability, NLL/CRPS, and separate epistemic from aleatoric behavior where feasible [35], [36].
Self-supervised / transfer	Improves label efficiency by reusing or pretraining representations.	Transfer learning has broader direct evidence [24], [33]; crop-specific self-supervised evidence remains limited but is growing [57], [75].	Negative transfer and overly optimistic random cross-validation remain major risks[25], [76].	Use cross-region, cross-year, cross-scale, spatially disjoint, low-shot, and zero-shot benchmarks with matched label budgets.
Physics-guided / hybrid	Injects agronomic structure to improve plausibility and extrapolation.	Coupled crop-model/ML and physics-guided studies report encouraging gains[58], [59].	Benefits can be confounded by richer inputs, tuning, or mis-specified priors.	Require matched baselines, transparent ablations, and explicit reporting of how physical guidance is

ARTICLE IN PRESS

	imposed [58], [35].

Taken together, these directions should not be collapsed into a single generic claim that they are 'promising.' They are promising for different reasons and at different levels of maturity. Uncertainty-aware models are methodologically essential but still under-evaluated for calibration; self-supervised and transfer-based approaches are strategically important for label scarcity but require stronger evidence under realistic transfer settings; and physics-guided models may offer the clearest route to extrapolative robustness, provided their added structure is validated against matched data-driven alternatives [35], [58], [59]. A longer-term opportunity is to combine these directions, for example, by using self-supervised remote-sensing encoders inside physics-guided predictors with calibrated probabilistic outputs. At present, however, that integration remains more a research agenda than an established benchmark. Following broader recommendations for machine-learning-based science, future crop-yield benchmarks should report not only point accuracy but also interval quality, cross-space and cross-time transfer, matched baseline and ablation analyses, and transparent release of code, pretrained weights, and data provenance, wherever feasible [45], [78].

5.8 Practical Guidance for Model Selection

Taken together, the literature supports a conditional hierarchy rather than a universal winner. Structured agronomic tables favor regularized MLPs and strong tabular baselines; satellite time series favor temporal CNNs, LSTMs, GRUs, or transformer-like models; UAV imagery favors CNN-based encoders when labels match the image scale; and multimodal problems justify hybrid or attention-based fusion only when the modalities contribute non-redundant information [5], [20]. Model selection should therefore be regime-first. Validation must prevent genotype, site, or trial leakage for tabular data; use spatial, temporal, or spatiotemporal holdouts for satellite sequences; control plot adjacency, repeated flights, and season effects for UAV imagery; and require alignment checks, ablations, and missing-modality analysis for multimodal systems. Table 7 summarizes these choices.

ARTICLE IN PRESS

Table 7: Practical decision guide for matching crop-yield data regimes to architecture families and validation precautions.

Data regime	Recommended architecture families	Main practical reason	Minimum validation precautions

ARTICLE IN PRESS

Structured genotype-environment-management tables	MLP/DNN, regularized feedforward models, tree-based baselines, tabular hybrids	Inputs are already summarized as agronomic descriptors, so strong tabular models can capture nonlinear interactions without unnecessary spatial or temporal complexity	Use genotype-, site-, environment-, or trial-level holdouts; avoid leakage from repeated genotypes, environments, or management records; compare against strong non-deep tabular baselines
Satellite time series and vegetation-index sequences	LSTM, GRU, temporal CNN, transformer-like encoders, CNN-LSTM hybrids	Yield signals are seasonal and spatially autocorrelated, so models need temporal structure and realistic transfer testing	Use spatial, temporal, or spatiotemporal splits; test unseen years or regions; control for cloud gaps, mixed pixels, administrative-label mismatch, and neighboring-location leakage
UAV, multispectral, hyperspectral, or proximal imagery	CNNs, 3D CNNs, ConvLSTM, CNN-attention models	High-resolution imagery contains local canopy, texture, spectral, and stress patterns that CNN-style encoders can extract effectively	Use plot-, field-, site-, or season-level holdouts; control for adjacent plots, repeated flights, illumination conditions, growth-stage differences, and small-sample overfitting
Multimodal fusion: EO, weather, soil, management, genotype	Hybrid CNN-RNN/MLP models, attention-based fusion, transformer-style multimodal models, physics-guided hybrids	Yield is shaped by interacting canopy, weather, soil, management, and genotype signals, so fusion is useful when modalities provide complementary information	Report modality alignment, missing-data handling, and ablation tests; use external or spatiotemporal validation; check whether gains persist when each modality is removed; evaluate uncertainty and missing-modality robustness

ARTICLE IN PRESS

What this hierarchy rejects is the idea that deeper or newer automatically means better. Architecture should be chosen by asking what structure in the data must be represented, what kinds

ARTICLE IN PRESS

of missingness and distribution shift are expected at deployment, and what levels of interpretability and computational burden are acceptable. Model selection in crop yield prediction is therefore a problem of evidence alignment, not architectural fashion [27], [58].

6. Explainability Techniques for Interpreting Models

6.1 Why Explanation Matters

Explanation is part of model evaluation, not a cosmetic afterthought. In crop yield prediction, models are rarely used solely to rank algorithms; they also support breeding, food security monitoring, insurance, and in-season management. Under those conditions, a prediction that cannot be interrogated is scientifically weak even when it is accurate. Recent evidence shows why this matters: explanation is needed to determine whether a model learned physiologically plausible relationships or merely exploited convenient correlations [79], [80], [81] while broader benchmarking work reminds us that apparently strong results can be artifacts of design choices rather than durable agronomic insight [35].

6.2 What Exactly Is Being Explained?

A county-level forecast, a field-scale estimate, a genotype-by-environment prediction, and a multimodal fusion model do not pose the same explanatory problem. In crop-yield studies, explanation may target a global ranking of predictors, a local case, a phenological window, a spatial hotspot, or a modality interaction. Failing to define this target leads to a common weakness in the literature: a single attractive heat map or feature ranking is presented as if the entire model has been explained. In reality, each explanation answers only a limited question, and its evidential value depends on whether that question matches the scientific or operational use case. Crop-yield studies already show both the promise and the limits of XAI. Paudel et al. [79] found that interpretable analyses of LSTM and 1D-CNN forecasts recovered plausible yield trends, static soil controls, and biomass signals, yet the same models were less successful in capturing extreme temperature and moisture effects. Explanations can therefore confirm that a model has learned a recognizable agronomic structure, but they can also reveal where the learned representation remains incomplete [79], [80].

6.3 Main Families of Explainability Methods

A more useful classification is not simply by algorithm name, but by what each method reveals and by the kind of error it risks. In crop-yield prediction, the same method can appear persuasive while answering the wrong question. Methods that rank variables, visualize spatial evidence, isolate important weeks, or explain a single county-year forecast should therefore be treated as complementary tools rather than interchangeable proofs of understanding.

6.3.1 Feature-Attribution Methods (SHAP, Integrated Gradients, and related Additive Scores).

Feature-attribution methods are useful in crop-yield prediction because they translate complex model behavior into local, global, temporal, and modality-level importance patterns. SHAP is particularly attractive because it can be applied across tabular, temporal, and mixed-input models, while Integrated Gradients and related additive scores provide complementary gradient-based views of how model inputs shape individual predictions. In crop-yield models, their value lies not only in ranking variables but also in assessing whether the model's importance pattern is consistent

ARTICLE IN PRESS

with crop phenology and known stress-sensitive growth stages. Recent wheat-yield studies illustrate this value. Guo et al. [51] used SHAP to interpret a transformer-based yield-estimation model. Li et al. [34] combined attention analysis with Integrated Gradients to identify critical periods associated with FPAR, LAI, and VTCI. Joshi et al. [82] used SHAP and Integrated Gradients to explain a Bi-LSTM model, showing that EVI, temperature, and precipitation during later growth stages strongly influenced predictions. Taken together, these studies show that attribution methods are most informative when variable importance is interpreted alongside phenological timing, modality relevance, and agronomic plausibility.

The main limitation is that attribution scores explain model credit assignment, not direct crop causality. Yield predictors such as LAI, FPAR, rainfall, temperature, vegetation indices, and drought indicators are often highly correlated, seasonally redundant, and partly proxy variables for unobserved management practices. Under these conditions, additive attribution methods can redistribute importance among related variables in ways that are mathematically coherent but scientifically ambiguous. A high attribution score should therefore be treated as evidence that the model used a signal, not as evidence that the signal independently caused the observed yield outcome. Accordingly, feature attribution should be presented as a diagnostic tool rather than a causal explanation. Stronger studies should report the attribution background or baseline choice, test stability across folds or random seeds, compare local and global explanations, and support attribution claims with perturbation, ablation, or stage-wise sensitivity checks. This protocol is especially important in multimodal crop-yield models, where several inputs may encode overlapping information about canopy development, water stress, and weather exposure.

6.3.2 Perturbation-Based Explanation (Permutation, Ablation, and Sensitivity Tests)

Perturbation-based explanation asks a harder question than attribution scores: Does predictive skill actually fall when a signal is disrupted? That is why it is especially valuable in crop-yield studies, where multiple variables encode overlapping phenological information. Khaki and Wang [5] showed the strong contribution of environmental variables in deep yield prediction, while more recent multimodal studies use modality removal and stage-wise sensitivity analysis to test whether satellite, weather, or soil data truly add independent information [25], [27]. The critical point, however, is that naive perturbation can generate unrealistic samples; permuting one month, one index, or one modality independently may break temporal dependence or biophysical coherence. A drop in accuracy is informative only when the perturbation itself respects agronomic structure.

6.3.3 Spatial Explanation for Image-Like Inputs (Saliency Maps and Grad-CAM).

For raster inputs or temporally stacked imagery, saliency maps and activation maps provide an intuitive picture of where the network is looking. Wolanin et al. [81] used regression activation maps to connect wheat yield estimates in the Indian Wheat Belt to growing-season length, temperature, and light conditions, showing that visual explanations can recover interpretable seasonal drivers. Yet this family of methods is also among the easiest to over-read. In crop-yield models, coarse saliency can highlight boundaries, background texture, cloud-contaminated regions, or sensor artifacts rather than crop processes. Grad-CAM-style maps are therefore most defensible when interpreted alongside crop calendars, spatial masks, and explicit sanity checks, rather than as standalone proof of biological relevance [83].

6.3.4 Intrinsic Explanation in Attention-Based Models.

ARTICLE IN PRESS

Attention is often treated as an appealing middle ground because it is built into the architecture rather than added after training. In crop-yield prediction, attention can identify influential weeks, variables, or modalities within sequence models and fusion networks. Tian et al. [50] showed attention concentrating on specific remotely sensed variables at key wheat stages; Aviles Toledo et al. [27] used attention within multimodal breeding models to emphasize influential feature groups; and Li et al. [34] and Najjar et al. [84] pushed this further by comparing temporal and modality attributions against crop phenology and Shapley-style baselines. This is useful evidence, but it is not conclusive. As the broader XAI literature cautions, attention weights are not automatically faithful explanations [54]. In a critical reading, attention is most persuasive when it is corroborated by perturbation tests, alternative attribution methods, or agronomic knowledge of growth stages.

6.3.5 Local Surrogate Explanation (LIME and Case-Level Diagnosis).

LIME is most helpful when the question is local: why did one county-year, field, or genotype receive a particular forecast? Joshi et al. [82] used LIME alongside SHAP and integrated gradients to interpret high-yield, low-yield, and high-error cases, and Pothapragada and Sujatha [38] used LIME and SHAP to make a maize-yield framework more communicable. The strength of LIME is communicative clarity at the instance level. Its weakness is that the local neighborhood around a crop-yield case can be unrealistic in high-dimensional agronomic space; synthetic perturbations may combine weather, vegetation, and management conditions that would never co-occur in practice. LIME should therefore be read as a diagnostic approximation of model behavior for a single case, not as a stable global account of what the model has learned.

6.3.6 Response Curve and Example-Based Analysis.

A different tradition asks whether the learned response surface itself is agronomically sensible. Hu et al. [80] showed that interpretable modeling can reveal nonlinear, asymmetric responses to temperature and precipitation and detect technological shifts that black-box models may obscure. This line of analysis is especially valuable for crop-yield prediction because it tests whether the learned relationship respects thresholds, saturation, and stress-response behavior, rather than merely fitting historical data. Its relative scarcity in deep-learning yield studies is itself a critical gap: a review should note not only which explanation tools are popular, but which agronomic questions remain underserved by current XAI practice.

6.4 Limits and Best Practices for Agronomic Explanation

Plausibility is not validation. A saliency map or SHAP ranking can look agronomically reasonable and still be unstable across seeds, folds, years, or neighboring regions. This problem is acute in crop-yield prediction because many dominant inputs, NDVI/EVI, LAI, FPAR, rainfall, temperature, and soil-water indicators, are correlated and seasonally structured. The literature, therefore, too often mistakes narrative plausibility for evidential strength. Omitted management and scale mismatch complicate interpretation. Yield is affected by cultivar choice, sowing date, fertilizer, irrigation, pests, and local management practices, but many datasets include only remote sensing, weather, and coarse soil variables. As a result, explanations may assign credit to proxies for mechanisms that are actually unobserved. The problem intensifies when field processes are explained using county-level labels or when field-scale heat maps are attached to aggregated yield targets. What appears to be a precise spatial explanation can therefore be a visually refined account of a coarse statistical association rather than a genuine causal signal [80], [81].

ARTICLE IN PRESS

Leakage and weak benchmark design can produce convincing but meaningless explanations. If training and test samples share neighboring fields, overlapping seasons, duplicated administrative units, or post-outcome information, the explanation may simply rationalize leakage. A model can then highlight important variables for the wrong reason, and the explanation may still look biologically sensible. This is why interpretability cannot be separated from validation design; leakage-aware splitting and deployment-relevant evaluation are prerequisites for trustworthy explanation [45], [35].

Transferability is an under-discussed limit of explanation. An attribution pattern learned in one region may not transfer to another where cultivars, management, disease pressure, or climate timing differ. Recent work on spatial transferability and domain adaptation in crop-yield prediction makes the same point about predictive accuracy; the interpretive implication is that explanations are also domain-bound unless tested across regions, years, or management contexts [33], [29].

Best practice should be protocol-driven. At minimum, studies should define the explanatory target, justify why the chosen method fits that target, report the background distribution or perturbation design, test stability across reruns or folds, compare explanations against phenological stages and agronomic expectations, and check whether attributions change when time windows or modalities are removed. The strongest recent crop-yield papers move in this direction, but the field as a whole still reports far more explanatory figures than explanatory validation [80], [81].

6.5 From Visualization to Validation

Local, global, and modality-level explanations should not be conflated. A local explanation explains why one prediction was high or low. A global explanation asks what the model learned across the dataset. A modality-level explanation explains how satellite, weather, soil, and management proxies were weighted in a fusion model. These are different evaluative tasks, yet many papers treat them as interchangeable outputs. A critical review should separate them because each requires different validation evidence [80], [82].

Explanation quality should be judged against multiple criteria. Fidelity asks whether the explanation aligns with the model's behavior. Stability asks whether it persists across seeds, folds, or nearby instances. Agronomic alignment assesses whether the highlighted signals align with crop physiology, stress timing, and expert knowledge. Transferability asks whether the same explanatory pattern survives across years or regions, while actionability asks whether the explanation helps users diagnose trust, uncertainty, or likely extrapolation. Current crop-yield studies still tend to focus more on plausibility than on stability or transfer [80], [82].

Uncertainty should be interpreted together with an explanation. Li et al. [34] showed that interpretability and uncertainty can be jointly analyzed, while the Bayesian neural-network work by Ma et al. [55] reminds us that yield predictions differ not only in the central estimate but also in confidence. An explanation attached to a highly uncertain forecast should not be treated the same as one attached to a well-calibrated prediction. For operational agriculture, the relevant question is not only what drove this forecast, but also how much confidence should be placed in that explanation under domain shift or sparse evidence.

The reviewed literature shows a clear distinction between explanation use and explanation validation. Many crop-yield studies use SHAP, LIME, saliency maps, Grad-CAM, attention weights, or activation maps mainly to visualize influential variables, time windows, or image

ARTICLE IN PRESS

regions. These outputs are valuable for interpretation, but they do not by themselves establish that an explanation is faithful, stable, or robust. A smaller subset moves closer to validation by comparing explanation methods, examining response behavior and high-error cases, testing perturbations or ablations, or checking agronomic plausibility. Table 8 maps representative studies according to this distinction.

Table 8: Evidence map of explanation use versus explanation validation in reviewed crop-yield studies.

Study	Explanation method used	What the study mainly did	Evidence of validation	Critical interpretation
[51]	SHAP	Interpreted important predictors in a transformer-based yield-estimation model	Partial; mainly feature-attribution interpretation	Useful for identifying model-used signals, but limited evidence that attributions were stable or causally meaningful
[79]	Interpretable analysis of LSTM and 1D-CNN forecasts	Related learned patterns to yield trends, soil controls, and biomass signals	Partial; agronomic plausibility was discussed	Stronger than visualization alone, but limited formal testing of explanation stability or perturbation robustness
[81],	Regression activation maps / spatial activation analysis	Linked spatial activation patterns to seasonal and environmental drivers of wheat yield	Partial; activation patterns were interpreted against environmental context	Helpful for spatial interpretation, but activation maps still require caution because they can highlight artifacts or correlated background structure

ARTICLE IN PRESS

ARTICLE IN PRESS

[82]	SHAP, LIME, and Integrated Gradients	Compared several explanation tools across representative wheat-yield cases	Partial to strong; multiple explanation methods and case-level analysis were used	More informative than single-method visualization, but stability across folds, seeds, or regions remains an important future requirement
[80]	Interpretable response-curve analysis	Examined whether learned climate-yield responses were agronomically sensible	Strong; response behavior was critically analyzed against known climate-yield relationships	Provides a stronger example of explanation as model diagnosis rather than decorative visualization
[84]	Attention attribution and Shapley-style comparison	Compared attention-derived relevance with alternative attribution evidence	Strong; explanations were compared rather than accepted at face value	Demonstrates a more defensible direction for validating attention-based explanations
[50]	Attention weights	Identified influential remote-sensing variables or growth stages	Partial; mainly attention-based interpretation	Useful for temporal interpretation, but attention weights alone should not be treated as faithful explanations
[38]	LIME and SHAP	Used local and global explanations to make maize-yield predictions more interpretable	Partial; mainly visualization and communicability	Improves transparency, but does not fully test fidelity, stability, or perturbation robustness

Partial validation means that the study went beyond simply reporting a prediction score or visual output, but did not fully test explanation fidelity, stability, perturbation robustness, and transferability. Strong means that the study compared explanation evidence, examined response behavior, or used explanation as part of model diagnosis rather than only as post hoc visualization.

ARTICLE IN PRESS

Based on this evidence, the literature supports a cautious conclusion. Existing explanation methods are useful for making crop-yield models more inspectable, but most studies still treat explanation as interpretation rather than as a validation object. The best-practice standard proposed in this review is therefore not merely to include SHAP values, LIME plots, Grad-CAM maps, saliency maps, or attention weights, but to test whether those explanations are faithful to the model, stable across data splits and random seeds, robust to perturbation or ablation, consistent with crop phenology and agronomic knowledge, and informative under high-error or high-uncertainty cases.

For crop-yield prediction, SHAP and LIME explanations should be checked quantitatively for stability across folds, random seeds, years, and regions; sensitivity to background-sample or perturbation choices; agreement with permutation importance, feature removal, or ablation results; and consistency between local and global explanations. Grad-CAM and saliency maps should be assessed using spatial sanity checks, deletion or insertion tests, crop-mask overlap, and robustness to cloud, shadow, soil-background, or non-crop pixels. Attention-based explanations should be compared with alternative attribution methods and tested through attention perturbation, time-window removal, or modality ablation. Ablation studies should report the size and variability of the resulting performance change, not merely whether a change occurred. Table 9 summarizes these method-specific validation checks.

Table 9: Quantitative validation checks for explanation methods in crop-yield prediction.

Explanation method	**Common visualization output**	**Recommended quantitative validation checks**
SHAP	Global or local feature-importance rankings	Test stability across folds, random seeds, regions, and years; assess sensitivity to background sample; compare with permutation importance, feature removal, or ablation results; compare local and global explanations
LIME	Case-level local explanation	Test stability under repeated perturbation sampling; assess sensitivity to neighborhood definition; compare with SHAP, Integrated Gradients, or feature-removal tests; inspect high-error and high-uncertainty cases
Grad-CAM and saliency maps	Spatial heatmaps over image-like inputs	Use deletion and insertion tests; check overlap with crop masks; test robustness to cloud, soil, shadow, and background pixels; compare explanations across growth stages, seasons, and sites
Attention weights	Time-window, feature, or modality relevance	Test attention perturbation; remove high-attention and low-attention time windows; perform modality ablation; compare attention patterns with SHAP, Integrated Gradients, or permutation importance
Ablation analysis	Performance change after removing features, time windows, model branches, or modalities	Report performance drop with confidence intervals or repeated-run variance; ensure removed components are agronomically meaningful; test across spatial, temporal, and spatiotemporal splits

ARTICLE IN PRESS

Explanation method	Common visualization output	Recommended quantitative validation checks
Agronomic plausibility checks	Alignment between explanations and known crop processes	Compare explanations with crop calendars, phenological stages, drought or heat-stress periods, management events, and expert agronomic expectations

Across all XAI methods, agronomic plausibility should be evaluated against crop calendars, known stress-sensitive growth stages, drought or heat-stress periods, and expert knowledge. High-error and high-uncertainty cases should also be inspected separately because explanations attached to unreliable predictions may be misleading even when they appear visually plausible. A practical validation standard for crop-yield XAI should therefore combine fidelity tests, stability analysis, perturbation or ablation checks, spatial sanity checks for image-based models, comparison across explanation methods, and agronomic plausibility assessment.

The explainability literature in CYP is therefore moving from descriptive visualization toward evidence-based model diagnosis. The strongest papers do not merely show what the model attended to; they demonstrate that the highlighted signals remain meaningful when subjected to validation tests. Yet even a well-explained model cannot be compared fairly if the benchmark itself is weak. The next methodological question, therefore, is not only whether a model can be interpreted, but whether competing models are being evaluated under protocols that reflect real agronomic deployment.

7. Benchmarking Requirements for Trustworthy Comparison

7.1 Split Design and Leakage Control

Benchmarking is where much of the apparent certainty in the crop-yield prediction literature breaks down. The most common reason published results appear stronger than they are is that the validation design fails to reflect the dependence structure of agricultural data. Nearby pixels, neighboring fields, repeated genotypes, overlapping weather histories, and repeated seasons can all leak information across training and test sets if splitting is done naively [45], [35], [76]. Random splits are, therefore, often overly optimistic in precisely those settings where authors later imply robustness under deployment.

External validation on unseen regions or years is especially important when a model is claimed to generalize under climate shift, management change, or geographic expansion. Benchmarks such as CropNet and the benchmark dataset for subnational crop-yield forecasting (CY-Bench) make split definitions more explicit and comparable [26], [85]. Even open datasets, however, do not automatically eliminate leakage: authors must still show that evaluation respects the true units of dependence in the data and the boundary between model development and deployment.

To make validation claims more transparent and actionable, split design should be matched to the type of generalization being claimed. Random splitting may be acceptable only for interpolation within already observed conditions; stronger claims require withholding the relevant spatial, temporal, genotypic, management, or operational context. Table 10 combines minimum validation requirements with common errors that can inflate performance, following the broader warning that leakage arises whenever the split fails to respect the true dependence structure of the data [45], [35], [76].

ARTICLE IN PRESS

Table 10: Minimum validation requirements and common validation errors for crop-yield prediction claims.

Claim type	What the claim implies	Minimum validation requirement	Common validation error to avoid	Leakage controls to report
Within-domain interpolation	The model performs well on new samples drawn from the same regions, seasons, genotypes, and management conditions represented in training	Random or stratified split may be acceptable, but should be repeated across seeds or folds	Reporting only one random split or allowing duplicated or near-identical records across train and test sets	Report repeated-run variance, stratification rules, duplicated records, repeated plots, and whether near-identical samples appear across train and test sets
Spatial transfer	The model generalizes to unseen fields, counties, farms, regions, or production zones	Spatial block, leave-location-out, leave-field-out, leave-county-out, or leave-region-out validation	Randomly splitting neighboring fields, plots, pixels, or counties across training and test sets	Report spatial distance or grouping rules; prevent neighboring plots, pixels, fields, or counties from appearing in both train and test sets; control for spatial autocorrelation
Temporal transfer	The model generalizes to future seasons or unseen time periods	Forward-chaining, leave-year-out, or train-past/test-future validation	Randomly splitting observations from the same year or using test-year information in preprocessing	Ensure that test-year information is not used in feature engineering, normalization, hyperparameter tuning, imputation, or augmentation
Cross-year generalization	The model remains reliable across different weather years, crop calendars, and seasonal anomalies	Leave-one-year-out or multi-year external testing, preferably including abnormal weather years	Reporting average performance across years without showing year-wise degradation or calibration in abnormal seasons	Report year-wise performance, calibration, and whether temporal summaries or climate normals include information from held-out years

ARTICLE IN PRESS

Claim type	What the claim implies	Minimum validation requirement	Common validation error to avoid	Leakage controls to report
Spatiotemporal transfer	The model generalizes to new places and new seasons simultaneously	Hold out region-year, site-year, field-year, or county-year combinations not observed during training	Holding out samples while still allowing the same region or same year to appear elsewhere in training	Prevent overlap in both geography and season; report whether the model was tuned on any part of the held-out region or year
Cross-genotype or breeding-program transfer	The model generalizes to unseen genotypes, hybrids, breeding lines, or trial environments	Leave-genotype-out, leave-family-out, leave-environment-out, or leave-trial-out validation	Allowing the same genotype, hybrid, pedigree, trial, or breeding environment in both training and test sets	Prevent repeated genotypes, pedigrees, trial identifiers, or closely related lines from leaking across splits unless the claim is explicitly within-family interpolation
Cross-scale deployment	The model trained at one scale remains useful at another scale, such as county-to-field or plot-to-farm prediction	External validation at the target deployment scale	Training and testing at mismatched scales without reporting aggregation effects or label-support differences	Report target definition, aggregation rules, spatial support, label provenance, and whether predictors and labels are aligned at the same scale
Operational deployment readiness	The model is suitable for real-world forecasting before harvest or under missing/noisy inputs	Prospective, external, or simulated real-time validation using only information available at the forecast date	Using post-harvest, future-season, cloud-free, complete, or manually cleaned inputs that would not be available operationally	Report forecast lead time, data latency, missing-modality handling, cloud-cover treatment, calibration, uncertainty, and failure cases

This is not a minor technicality. In county-scale maize and soybean forecasting, adjacent Corn Belt counties may share climate trajectories, crop calendars, and management patterns; in breeding datasets, the same genotypes or trial structures can recur across environments; and in UAV studies, overlapping flights and similar illumination conditions can make holdout samples far less independent than they appear [22], [30]. A trustworthy evaluation must specify whether splits are random, spatial, temporal, spatiotemporal, or hierarchical, and it should justify that choice relative

ARTICLE IN PRESS

to the intended use case. Spatial transfer, future-year forecasting, and new-environment deployment require harder splits than interpolation within familiar conditions [30], [35], [76].

7.2 Metrics, Calibration, Uncertainty, and Baseline Comparison

Benchmark reporting should go beyond a single accuracy value. RMSE, MAE, bias, and correlation-based metrics such as R or R^2 remain useful, but they should be interpreted in relation to target scale, forecast horizon, crop type, and unit of analysis. For example, a five-percent error at the subnational scale late in the season is not equivalent to the same error at the field scale two months before harvest. Without this contextualization, cross-study comparison becomes numerically tidy but scientifically weak [34], [35]. Baseline comparison is equally important. Many papers compare deep models only against weak or outdated baselines, which can overstate the apparent value of the neural architecture. Strong benchmarking should include naive seasonal baselines, linear or tree-based comparators, and, where appropriate, agronomic or process-informed references. The right question is not whether deep learning beats a weak comparator, but whether it adds sufficient value over credible alternatives to justify its cost and complexity.

Across the reviewed benchmarking literature, the main finding is that validation rigor is uneven rather than absent. Several benchmark-oriented or spatial-transfer studies report spatial, temporal, or external tests, but many architecture-focused studies still rely on random or weakly described splits that may not support claims of spatial transfer, temporal transfer, or deployment readiness. Similarly, most studies report point-accuracy metrics such as RMSE, MAE, or R^2, whereas fewer report calibration, interval quality, repeated-run variance, or uncertainty behavior under domain shift. The recommendation advanced in this review is therefore stronger than a description of current practice: future studies should match the validation split to the generalization claim, compare against credible non-deep and simple neural baselines, report uncertainty calibration when probabilistic forecasts are produced, and disclose enough preprocessing, split, and tuning detail to make the benchmark auditable.

Calibration and uncertainty remain major omissions. A model may be accurate on average yet become overconfident when observations are missing, during extreme weather, or outside the training domain. Probabilistic models should therefore report empirical coverage, interval sharpness, reliability diagnostics, and proper scoring rules such as negative log-likelihood, continuous ranked probability score, Brier score, interval score, or energy score [55], [86].Trustworthy deployment also requires safe-AI checks, including subgroup performance, robustness to degraded inputs, calibration under distribution shift, and warning or abstention behavior under high uncertainty[87]. Table 11 summarizes these metric categories.

Table 11: Metric categories for trustworthy crop-yield prediction.

Metric category	Examples	Purpose
Point accuracy	RMSE, MAE, bias, R^2	Measures average prediction error, but does not assess uncertainty reliability
Proper scoring rules	Negative log-likelihood, continuous ranked probability score, Brier score, interval score, energy score	Evaluates probabilistic forecasts, calibration, and sharpness

ARTICLE IN PRESS

Calibration	Reliability diagrams, empirical coverage, calibration curves	Tests whether predicted probabilities or intervals match observed outcomes
Safe-AI and robustness metrics	Subgroup error, worst-group error, missing-modality robustness, out-of-distribution performance, abstention rate	Assesses whether models remain reliable under operational stress and deployment shift
Reproducibility metrics	Repeated-run variance, confidence intervals, seed sensitivity	Shows whether reported performance is stable and auditable

Future benchmarks should therefore combine point-accuracy metrics with proper scoring rules, calibration diagnostics, robustness checks, and safe-AI-oriented indicators of failure behavior. Because agricultural risk is often concentrated in rare low-yield seasons, benchmark reports should also include tail-sensitive evaluation, such as quantile loss, exceedance-probability calibration, low-yield-event detection, Value-at-Risk, and Expected Shortfall, rather than relying only on average-error metrics.

7.3 Transferability: Why the Maize/Corn example matters

The maize/corn literature provides a useful stress test for benchmarking because it combines strong public data availability with high practical relevance. U. S. Corn Belt studies have produced influential deep-learning results using county-level remote sensing, hybrid architectures, transfer learning, and Bayesian uncertainty analysis [6], [24]. Their value lies not in implying that maize is a solved problem, but in showing how sensitive performance is to benchmark context. The Corn Belt offers relatively dense yield statistics, stable remote-sensing coverage, and large contiguous production zones, all of which are favorable for data-driven learning compared with fragmented smallholder systems or sparsely monitored regions [88], [55].

This creates a transferability risk. A model that works well on county-level Corn Belt data may rely on climatic regularities, planting patterns, and statistical smoothness that are absent in other crops, countries, or finer-resolution field settings. Small-field transfer studies and domain-adaptation work demonstrate that performance can degrade when landscape structure, management heterogeneity, or sensor context changes [33], [76]. The maize/corn example, therefore, matters precisely because it is often mistaken for a universal benchmark when, in fact, it is a best-case environment for many current methods. Readers should treat such results as evidence of potential under data-rich conditions, not as proof of broad crop-agnostic deployability.

7.4 Reproducibility, Comparability, and Reporting Artifacts

A trustworthy benchmark is defined not only by its dataset and score, but also by the transparency of the workflow that links raw inputs to published claims. In crop yield prediction, reproducibility is often limited because studies do not report critical methodological details, such as pre-processing pipelines, season-selection rules, cloud-masking procedures, feature engineering choices, split logic, hyperparameter-tuning budgets, and random seeds [45], [35]. As a result, studies that appear comparable may actually rely on different assumptions, data treatments, or validation conditions, which weakens both replication and fair comparison [35], [78].

ARTICLE IN PRESS

This problem is even more acute in multimodal systems because small implementation decisions can produce large apparent gains. Choices about alignment windows, imputation rules, resampling frequency, missing-modality handling, and augmentation policies can effectively redefine the task itself. Therefore, a robust benchmark should report how each modality is aligned, cleaned, resampled, and fused, so readers can distinguish genuine model improvements from pre-processing effects or benchmark artifacts. Model cards and data cards provide useful reporting templates by documenting intended use, limitations, provenance, and evaluation design in a consistent and auditable way [78], [89].

A fair comparison also depends on explicit disclosure of the computational budget and the tuning protocol. A transformer explored over a large hyperparameter space should not be presented as superior to a simpler baseline that received only minimal tuning. For this reason, benchmark reports should include the search strategy, number of runs, variance across runs, early-stopping criteria, hardware context, training time, and parameter counts. These details are especially important in crop-yield prediction, where many reported improvements are modest and may disappear when evaluation conditions are more rigorously matched [35], [29]. Taken together, these reporting artifacts should be treated as part of the benchmark itself rather than as supplementary detail.

7.5 Open Benchmark Ecosystems and Their Limits

Public benchmark ecosystems are beginning to improve comparability. CropNet provides a large-scale, multimodal, climate-aware dataset for county-level U. S. forecasting [23], while CY-Bench extends benchmarking to subnational yield forecasting across multiple countries and production settings [68]. These resources encourage clearer split definitions, stronger external comparison, and more cumulative progress. Their limits, however, should be stated plainly: CropNet remains anchored to the U. S. county structure, while CY-Bench is constrained by the crops, administrative units, and predictor sets that can be assembled consistently at the subnational scale [24], [68]. Open benchmarks are therefore enabling infrastructure rather than complete representations of agricultural reality. Tables 12 and 13 summarize the information needed to make benchmark claims interpretable, comparable, and reproducible [33], [43], [59], [61], [71].

Table 12: Minimum reporting expectations for targets, data, validation, and metrics.

Benchmark component	Minimum reporting expectation	Why it matters
Target definition	Report the crop, spatial unit, unit of measurement, aggregation level, and forecast lead time	Error metrics are not comparable when the target itself changes
Split strategy	Specify whether splits are random, spatial, temporal, spatiotemporal, or hierarchical, and disclose leakage and overlap controls	Prevents optimistic leakage and clarifies deployment realism
Input modalities and alignment	Describe modality timing, spatial resolution, missingness handling, and fusion strategy	Performance may depend more on alignment choices than on model family

ARTICLE IN PRESS

Metrics, uncertainty, and calibration	Report RMSE, MAE, bias, and R^2 at a minimum; include calibration when probabilistic predictions are used; and state the forecast horizon to which each metric applies	Supports risk-sensitive interpretation rather than point accuracy alone
Baselines and variance reporting	Report naive, statistical, or agronomic baselines, repeated-run variance when stochastic training is used, and confidence intervals or significance tests when superiority is claimed	Distinguishes meaningful gains from noise or weak comparators
External validation	Test on unseen years, regions, sites, or production environments whenever transportability is claimed	Determines whether results survive a domain shift

Table 13: Minimum reporting expectations for explainability, reproducibility, and computational transparency.

Benchmark component	**Minimum reporting expectation**	**Why it matters**
Explainability validation	State the explanation target and evaluate stability, fidelity, or perturbation robustness	Distinguishes faithful explanation from plausible visualization
Reproducibility artifacts	Release code, seeds, pre-processing scripts, split logic, feature-engineering rules, and key configuration details when licensing allows	Makes comparison cumulative and auditable
Computation and tuning budget	State the hyperparameter-search budget, early-stopping policy, training repeats, software libraries, and hardware used to obtain the reported results	Prevents unfair comparisons driven by unequal optimization effort rather than genuine methodological advantage

ARTICLE IN PRESS

To make the methodological limitations more visible, Table 14 maps major reviewed studies against key evidence-quality indicators. The table does not rank studies by predictive accuracy. Instead, it records whether each study reports validation practices needed to support trustworthy comparison, including spatial or temporal splitting, external validation, strong baseline comparison, uncertainty calibration, code or data availability, and validation of explanation methods. This evidence map shows that many reported advances remain difficult to compare because methodological safeguards are unevenly reported across model families, data regimes, and prediction settings.

Table 14: Evidence map of methodological safeguards reported in major crop-yield prediction studies.

ARTICLE IN PRESS

Study	Data regime / setting	Model family	Spatial or temporal split reported	External validation	Strong baseline comparison	Uncertainty calibration	Code/data availability	Explanation validation
[5]	Structured genotype-environment-management tables	DNN / MLP	P	P	Y	N	P	N
[7]	Satellite imagery and climate data	CNN / deep Gaussian process	P	P	Y	P	P	NR
[24]	Satellite time series; corn and soybean	CNN-LSTM / transfer learning	P	P	Y	N	P	N
[20]	UAV multimodal imagery	CNN / multimodal learning	P	P	Y	N	P	N
[30]	Field-scale / spatial yield mapping	Spatial ML / DL comparison	Y	P	Y	N	P	NR

ARTICLE IN PRESS

ARTICLE IN PRESS

[31]	Spatially heterogeneous crop-yield mapping	Spatial DL / ML models	Y	P	Y	N	P	NR
[27]	Multimodal breeding / management data	Attention-based multimodal model	P	P	Y	N	P	P
[33]	Sub-field or cross-domain transfer setting	Domain adaptation / multimodal DL	Y	Y	Y	N	P	NR
[55]	Corn-yield uncertainty analysis	Bayesian neural network	P	P	Y	Y	P	NR
[81]	Remote-sensing wheat-yield prediction	CNN / activation maps	P	P	Y	N	P	P
[79]	Large-scale crop-yield forecasting	LSTM / 1D-CNN with XAI	P	P	Y	N	P	P
[80]	Climate-yield	Interpretable ML	P	P	Y	N	P	Y

ARTICLE IN PRESS

ARTICLE IN PRESS

	response analysis							
[51]	Wheat-yield estimation	Transformer / SHAP	P	P	Y	N	P	P
[82]	Wheat-yield prediction	Bi-LSTM with SHAP, LIME, IG	P	P	Y	N	P	P
[84]	Attention and attribution analysis	Attention-based model / Shapley comparison	P	P	Y	N	P	Y
[26]	County-scale multimodal benchmark	Benchmark dataset / multimodal models	Y	P	Y	P	Y	NR
[85]	Subnational crop-yield benchmark	Benchmark dataset / forecasting models	Y	Y	Y	P	Y	NR

Y = clearly reported; P = partially reported or limited; N = not reported; NR = not relevant to the study objective. “Strong baseline comparison” refers to comparison with credible non-deep, statistical, agronomic, or simpler neural baselines rather than only weak or outdated comparators. “Explanation validation” refers to stability, fidelity, perturbation, ablation, or comparison across explanation methods, rather than visualization alone.

ARTICLE IN PRESS

The evidence map supports three conclusions. First, spatially or temporally explicit validation is becoming more common, especially in benchmark-oriented and transfer-focused studies, but random or weakly described splitting remains frequent in architecture-focused papers. Second, uncertainty calibration and explanation validation are still much less consistently reported than point-accuracy metrics, even when models are intended for risk-sensitive or interpretable agricultural decision-making. Third, code, data, split definitions, and preprocessing details remain unevenly available, which limits reproducibility and makes it difficult to determine whether reported gains arise from model architecture, data regime, preprocessing choices, or validation design.

8. Discussion

The preceding sections point to a broader methodological conclusion: in crop-yield prediction, the central question is not whether deep learning can be accurate, but under what evidential conditions its predictions deserve trust. The literature already shows that deep learning can be useful in breeding, remote sensing, and multimodal forecasting settings [5], [27]. What remains unresolved is how much of the reported improvement reflects genuine learning of agronomically meaningful structure and how much reflects favorable data regimes, permissive validation, or benchmark-specific engineering.

From this perspective, dataset choice is not a background detail but a core component of evaluation. Structured breeding records, county-level public archives, UAV phenotyping datasets, and multimodal operational benchmarks reward different inductive biases and support different levels of inference. For that reason, the question "Which model is best?" is less informative than "Which model is most defensible for this data regime, target, and decision context?" Once evaluation is framed this way, several apparent contradictions in the literature become easier to interpret: feedforward networks can remain competitive, complex hybrids can be justified, and transformer-based models can be promising without being uniformly superior [20], [88].

A second implication concerns the link between explainability and benchmarking. Explanations are valuable only when the underlying benchmark establishes that the model itself is valid. A faithful SHAP analysis of a leakage-prone model still interprets a flawed evidential base, whereas a strong external benchmark without explanation still leaves the agronomic meaning of the prediction underspecified. Trustworthy deployment, therefore, requires data quality, model choice, explanation, and validation protocol to be presented as parts of a single evidential argument rather than as disconnected checklist items. Finally, public ecosystems such as CropNet and CY-Bench lower the barrier to rigorous comparison, but they do not eliminate the need for harder external tests, richer management and genotype data, or decision-centered evaluation [26], [35], [85].

8.1 Operational Deployment Barriers in Applied Digital Agriculture

Although much of the crop-yield prediction literature evaluates models under offline benchmark conditions, operational deployment in digital agriculture introduces additional constraints that are not captured by accuracy metrics alone. A model that performs well retrospectively may still be difficult to use in real farm, advisory, insurance, or food-security systems if required input streams arrive late, are incomplete, or cannot be integrated into existing decision workflows.

ARTICLE IN PRESS

One major barrier is data latency. Satellite observations may be delayed by revisit cycles, cloud contamination, preprocessing pipelines, or product-release schedules, while weather, soil, management, and yield-label data may arrive at different temporal resolutions and with different levels of quality control. These delays are especially important for in-season forecasting because the practical value of a prediction depends not only on its accuracy, but also on whether it is available early enough to support irrigation, fertilizer, harvest, logistics, insurance, or market decisions. Label latency is equally problematic: harvested yield records are often available only after the season, may require cleaning or moisture correction, and may be aggregated at administrative scales that are poorly aligned with farm-level decisions.

A second barrier is missing or unreliable modalities. In operational settings, remote-sensing inputs can be affected by cloud cover, sensor gaps, atmospheric correction errors, and inconsistent revisit frequency. Farm-management data may be missing, manually recorded, proprietary, or stored in formats that are not easily linked to remote-sensing and weather data. Multimodal models are therefore useful only if they remain robust when one or more inputs are delayed, noisy, or unavailable. For applied IoT and digital agriculture contexts, this implies that models should report missing-modality performance, fallback behavior, imputation rules, and uncertainty changes under degraded input conditions.

A third barrier is real-time or near-real-time inference. Large hybrid, transformer-based, or multimodal models may be computationally demanding, especially when predictions must be updated frequently across many fields or regions. Operational systems may require lightweight architectures, model compression, edge-cloud partitioning, or staged inference pipelines in which low-cost models provide preliminary estimates and more complex models are used only when sufficient data are available. Reporting training accuracy without inference cost, update frequency, hardware assumptions, and scalability therefore gives an incomplete picture of deployment readiness.

These deployment barriers are especially severe in low-resource agricultural settings and many developing-country contexts. Yield labels may be sparse, delayed, aggregated, or inconsistently recorded; farm-management data are often informal or unavailable; internet connectivity, cloud-computing access, and local sensor infrastructure may be limited; and smallholder landscapes can contain heterogeneous fields that are poorly represented by coarse administrative or satellite products. Under these conditions, technically sophisticated models may fail not because the architecture is weak, but because the operational data pipeline cannot support the required inputs at the right spatial resolution, temporal frequency, or quality level. Deployment-oriented crop-yield models for such settings should therefore prioritize data-efficient learning, uncertainty reporting, missing-modality robustness, lightweight inference, locally calibrated baselines, and integration with extension services or decision-support systems that match local institutional capacity.

Finally, deployment depends on integration with farm management systems and decision-support workflows. Yield forecasts must be delivered in formats that farmers, extension workers, insurers, or policy analysts can interpret and act upon. This requires interoperability with farm-management information systems, IoT sensor platforms, geospatial dashboards, advisory tools, and enterprise planning systems. It also requires clear communication of uncertainty, forecast horizon, spatial

ARTICLE IN PRESS

resolution, and known failure modes. Without such integration, even technically strong models may remain research prototypes rather than operational tools for digital agriculture.

8.2 Challenges and Limitations

Several structural challenges still limit the reliability and transferability of deep-learning-based crop-yield prediction. The first is multimodal incompleteness. Yield is jointly shaped by genotype, weather, soil, management, pests, and local disturbances, yet most datasets observe only a subset of these factors. Even large public benchmarks often omit management detail, cultivar information, or field history, forcing models to infer key drivers from imperfect proxies [26], [35], [36]. The consequence is not only lower accuracy, but also a persistent risk that models learn unstable correlations that fail under new conditions.

A second challenge is targeting scarcity and inconsistency. Harvested yield labels are often delayed, aggregated administratively, measured under different moisture corrections, or unavailable at the spatial scale implied by the inputs. As a result, some of the field's apparent methodological progress is built on targets whose comparability is weaker than the reported precision of the evaluation metrics suggests. Small, high-quality breeding or multimodal datasets and large, noisier public datasets each address different parts of the problem, yet the literature still lacks a clear strategy for linking them [20], [21], [35].

A third challenge is distribution shift. Models are often optimized for interpolation within familiar seasons, regions, or management systems, whereas deployment requires credible performance under new cultivars, extreme weather, different farming systems, or new geographies. Spatial-transfer studies and uncertainty-aware analyses both suggest that performance can deteriorate sharply under changed conditions, even when in-domain accuracy appears strong [29], [33]. Put differently, much of the literature still measures convenience generalization rather than operational generalization.

This review also has important limitations. Because the literature is heterogeneous in crop, scale, metric, and evaluation protocol, the synthesis is interpretive rather than meta-analytic. The review is structured but not registered, and it does not provide a pooled quantitative effect estimate or a database-by-database audit trail. It also focuses on deep learning rather than offering a full comparison with process-based models or the complete range of classical statistical approaches. Finally, the analysis depends on what published studies report about data splits, pre-processing, uncertainty, and tuning. These constraints reinforce rather than weaken the core conclusion: the field needs better reporting and greater benchmark transparency precisely because rigorous evaluation remains limited by under-specified methods.

8.3 Future Directions

Future work should prioritize directions that improve realism, transferability, and transparency together, rather than pursuing marginal gains on isolated benchmarks. The first priority is benchmark-centered progress. Open datasets such as CropNet and CY-Bench provide an important foundation, but they should be extended with harder spatiotemporal splits, stronger baseline suites, management-aware subsets, and standardized reporting of calibration, variance, and external transfer [26], [35].

A second priority is label efficiency and representation learning. Satellite, meteorological, and environmental archives are far richer than harvested-yield labels, so self-supervised pretraining

ARTICLE IN PRESS

and contrastive learning are promising approaches for learning transferable spatiotemporal structure before fine-tuning on scarce targets [57], [75]. Transfer-learning results across crops further suggest that pretraining may be especially valuable when the target crop or region has limited labelled data [24]. The critical question for future work is whether such pretraining improves out-of-domain performance rather than merely increasing in-domain efficiency.

A third priority is extrapolation under climate and management shift. Physics-guided or knowledge-informed deep learning is attractive because it can constrain predictions with agronomic structure when historical interpolation is no longer sufficient. Uncertainty-aware models are equally important because agricultural decisions are made under risk, not under average-case conditions. These directions should converge: future systems should report not only what they predict, but also how confidently they predict under shift and which mechanistic assumptions support that confidence [59], [33].

A fourth priority is to treat explanation as evaluation rather than ornament. Multimodal and attention-based systems will continue to expand, but their explanations should be tested through perturbation, ablation, and stability analyses. Model cards and data cards should also become routine in crop-yield prediction, because they force authors to state intended use, benchmark limits, and known failure modes more explicitly [78], [80].

A further priority is tail-risk modeling for rare but high-impact agricultural events. Most crop-yield prediction benchmarks emphasize average predictive performance, yet agricultural decision-makers are often most concerned with droughts, heatwaves, floods, and severe yield-loss years. Extreme Value Theory (EVT) and related tail-risk approaches provide a useful direction for modeling the probability and severity of rare outcomes that may be poorly represented in standard training data. In crop-yield prediction, this implies evaluating not only mean error, but also performance in the lower tail of the yield distribution, including severe underproduction events, crop-failure risk, and compound climate stresses. Future benchmarks should therefore report tail-sensitive metrics such as low-yield recall, quantile loss for extreme quantiles, exceedance-probability calibration, Value-at-Risk, Expected Shortfall, and event-based Brier scores for yield-loss thresholds. This would make trustworthy crop-yield prediction more relevant to food-security planning, crop insurance, and climate-risk management, where rare failures can matter more than average-season accuracy [77].

Finally, following broader recommendations for machine-learning-based science, future crop-yield benchmarks should report not only point accuracy, but also interval quality, cross-space and cross-time transfer, matched baseline and ablation analyses, and transparent release of code, pretrained weights, and data provenance wherever feasible [45], [89]. Taken together, these priorities imply a shift in research culture: the most valuable next-generation crop-yield models will be those that remain reliable when data are scarce, labels are noisy, modalities are missing, and deployment conditions differ from the training benchmark.

9. Conclusion

Deep learning has expanded what is technically possible in crop-yield prediction, but current evidence does not support a single architecture as universally superior across crops, scales, or data regimes. The stronger conclusion is that performance depends on model-data fit: structured agronomic tables, remote-sensing imagery, seasonal sequences, and multimodal operational pipelines reward different inductive biases and should not be treated as interchangeable tasks. This

ARTICLE IN PRESS

review is therefore methodological rather than celebratory. Trustworthy crop-yield prediction depends on five linked requirements: a clearly defined target, a transparent review or benchmark protocol, architecture choices matched to the data regime, explanations validated for fidelity and stability, and evaluations designed to resist leakage and test transportability. When any link in this chain is weak, headline accuracy is a poor proxy for real-world usefulness. When the chain is strong, even modest gains can be scientifically and operationally meaningful. That is the evaluation standard the field should adopt if deep learning is to support robust, transparent, and trustworthy agricultural decision-making.

Declarations

Authors' contributions

Danison Taremwa: Conceptualization, methodology, investigation, and original draft. Emmanuel Ahishakiye: Supervision, review, and editing. Aggrey Obbo: Review and editing. Paul Kategaya Kisozi: Review and editing. Fred Kaggwa: Supervision, review, and editing. All authors read and approved the final manuscript.

Funding statement

The authors received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgments

None.

Competing interests

The authors declare that they have no competing interests.

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent to publish

Not applicable.

Data availability statement

Not applicable.

Code availability statement

No custom code was generated for this review article.

REFERENCES

ARTICLE IN PRESS

ARTICLE IN PRESS

[1] S. B. Khadija Meghraoui , Imane Sebari , Juergen Pilz, Kenza Ait El Kadi, “Applied Deep Learning-Based Crop Yield Prediction : A Systematic Analysis of Current Developments and Potential Challenges,” 2024.

[2] L. E. I. Liu and X. Song, “Accurate crop yield prediction via temporal convolutional network and variational autoencoder,” vol. 49, no. 3, pp. 612–623, 2025, doi: 10.55730/1300-011X.3290.

[3] T. van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” *Comput. Electron. Agric.*, vol. 177, no. January, p. 105709, 2020, doi: 10.1016/j.compag.2020.105709.

[4] P. Muruganantham, S. Wibowo, S. Grandhi, N. H. Samrat, and N. Islam, “A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing,” *Remote Sens.*, vol. 14, no. 9, 2022, doi: 10.3390/rs14091990.

[5] S. Khaki and L. Wang, “Crop yield prediction using deep neural networks,” *Front. Plant Sci.*, vol. 10, no. May, pp. 1–10, 2019, doi: 10.3389/fpls.2019.00621.

[6] S. Khaki, L. Wang, and S. V Archontoulis, “A CNN-RNN Framework for Crop Yield Prediction,” vol. 10, no. January, pp. 1–14, 2020, doi: 10.3389/fpls.2019.01750.

[7] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, “Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data,” pp. 4559–4565, 2014.

[8] A. Oikonomidis, C. Catal, and A. Kassahun, “Deep learning for crop yield prediction: a systematic literature review,” *New Zeal. J. Crop Hortic. Sci.*, vol. 0, no. 0, pp. 1–26, 2022, doi: 10.1080/01140671.2022.2032213.

[9] R. Wang *et al.*, “Transformers for Remote Sensing: A Systematic Review and Analysis,” *Sensors*, vol. 24, no. 11, 2024, doi: 10.3390/s24113495.

[10] Y. Özüpak and E. Aslan, “An enhanced interpretable ML approach for forecasting long-term CO2 and N2O emissions via optimized random forest modeling,” *Fuel*, vol. 427, p. 139723, 2027, doi: https://doi.org/10.1016/j.fuel.2026.139723.

[11] J. Shen *et al.*, “Finger Vein Recognition Algorithm Based on Lightweight Deep Convolutional Neural Network,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022, doi: 10.1109/TIM.2021.3132332.

[12] J. Shen, W. Zhou, N. Liu, H. Sun, D. Li, and Y. Zhang, “An Anchor-Free Lightweight Deep Convolutional Network for Vehicle Detection in Aerial Images,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24330–24342, 2022, doi: 10.1109/TITS.2022.3203715.

[13] S. Archana and P. Senthil Kumar, “A Survey on Deep Learning Based Crop Yield Prediction,” *Nat. Environ. Pollut. Technol.*, vol. 22, no. 2, pp. 579–592, 2023, doi: 10.46488/NEPT.2023.v22i02.004.

[14] A. Jabed, M. Azrifah, and A. Murad, “Heliyon 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, p. e40836, 2024, doi: 10.1016/j.heliyon.2024.e40836.

ARTICLE IN PRESS

[15] A. Oikonomidis, C. Catal, and A. Kassahun, "Hybrid Deep Learning-based Models for Crop Yield Prediction," *Appl. Artif. Intell.*, vol. 36, no. 1, 2022, doi: 10.1080/08839514.2022.2031823.

[16] A. Rajbongshi *et al.*, "Leveraging explainable AI for sustainable agriculture: a comprehensive review of recent advances," *Artif. Intell. Rev.*, vol. 59, no. 3, p. 105, 2026, doi: 10.1007/s10462-025-11459-5.

[17] Z. Liu *et al.*, "International Journal of Applied Earth Observation and Geoinformation In-season crop yield prediction : State of the art and future research direction," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 146, no. December 2025, p. 105129, 2026, doi: 10.1016/j.jag.2026.105129.

[18] H. Snyder, "Literature review as a research methodology : An overview and guidelines," *J. Bus. Res.*, vol. 104, no. July, pp. 333–339, 2019, doi: 10.1016/j.jbusres.2019.07.039.

[19] M. J. Page *et al.*, "The PRISMA 2020 statement : an updated guideline for reporting systematic reviews Systematic reviews and Meta-Analyses," 2021, doi: 10.1136/bmj.n71.

[20] M. Maimaitijiang, V. Sagan, P. Sidike, S. Hartling, F. Esposito, and F. B. Fritschi, *Soybean yield prediction from UAV using multimodal data fusion and deep learning*, vol. 237, no. 1. 2020. doi: 10.1016/j.rse.2019.111599.

[21] P. Nevavuori, N. Narra, P. Linna, and T. Lipping, "Crop yield prediction using multitemporal UAV data and spatio-temporal deep learning models," *Remote Sens.*, vol. 12, no. 23, pp. 1–18, 2020, doi: 10.3390/rs12234000.

[22] J. Shook, T. Gangopadhyay, L. Wu, B. Ganapathysubramanian, S. Sarkar, and A. K. Singh, "Crop yield prediction integrating genotype and weather variables using deep learning," *PLoS One*, vol. 16, no. 6 June 2021, pp. 1–19, 2021, doi: 10.1371/journal.pone.0252402.

[23] J. Sun, L. Di, Z. Sun, Y. Shen, and Z. Lai, "County-level soybean yield prediction using deep CNN-LSTM model," *Sensors (Switzerland)*, vol. 19, no. 20, pp. 1–21, 2019, doi: 10.3390/s19204363.

[24] S. Khaki, H. Pham, and L. Wang, "Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning," *Sci. Rep.*, vol. 11, no. 1, pp. 1–14, 2021, doi: 10.1038/s41598-021-89779-z.

[25] F. Mena *et al.*, "Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction," *Remote Sens. Environ.*, vol. 318, p. 114547, 2025, doi: https://doi.org/10.1016/j.rse.2024.114547.

[26] F. Lin, K. Guillot, S. Crawford, Y. Zhang, X. Yuan, and N. F. Tzeng, "An Open and Large-Scale Dataset for Multi-Modal Climate Change-aware Crop Yield Predictions," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, no. April, pp. 5375–5386, 2024, doi: 10.1145/3637528.3671536.

[27] C. Aviles Toledo, M. M. Crawford, and M. R. Tuinstra, "Integrating multi-modal remote sensing, deep learning, and attention mechanisms for yield prediction in plant breeding experiments," *Front. Plant Sci.*, vol. 15, no. July, pp. 1–16, 2024, doi: 10.3389/fpls.2024.1408047.

ARTICLE IN PRESS

[28] J. Cao *et al.*, “Mapping global yields of four major crops at 5-minute resolution from 1982 to 2015 using multi-source data and machine learning,” *Sci. Data* , vol. 12, no. 1, pp. 1–15, 2025, doi: 10.1038/s41597-025-04650-4.

[29] S. Stiller, K. Grahmann, G. Ghazaryan, and M. Ryo, “Improving spatial transferability of deep learning models for small-field crop yield prediction,” *ISPRS Open J. Photogramm. Remote Sens.*, vol. 12, p. 100064, 2024, doi: https://doi.org/10.1016/j.ophoto.2024.100064.

[30] P. Killeen, I. Kiringa, T. Yeap, and P. Branco, “Corn Grain Yield Prediction Using UAV-Based High Spatiotemporal Resolution Imagery , Machine Learning , and Spatial Cross-Validation,” pp. 1–39, 2024.

[31] S. Zhou, L. Xu, and N. Chen, “Rice Yield Prediction in Hubei Province Based on Deep Learning and the Effect of Spatial Heterogeneity,” *Remote Sens.*, vol. 15, no. 5, pp. 1–17, 2023, doi: 10.3390/rs15051361.

[32] M. S. Mia *et al.*, “Multimodal Deep Learning for Rice Yield Prediction Using UAV-Based Multispectral Imagery and Weather Data,” *Remote Sens.*, vol. 15, no. 10, 2023, doi: 10.3390/rs15102511.

[33] Y. Ma, Z. Yang, Q. Huang, and Z. Zhang, “Improving the Transferability of Deep Learning Models for Crop Yield Prediction: A Partial Domain Adaptation Approach,” *Remote Sens.*, vol. 15, no. 18, p. 4562, 2023, doi: 10.3390/rs15184562.

[34] M. Li *et al.*, “An interpretable wheat yield estimation model using an attention mechanism-based deep learning framework with multiple remotely sensed variables,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 140, p. 104579, 2025, doi: https://doi.org/10.1016/j.jag.2025.104579.

[35] P. Filippi, S. Y. Han, and T. F. A. Bishop, “On crop yield modelling, predicting, and forecasting and addressing the common issues in published studies,” *Precis. Agric.*, vol. 26, no. 1, p. 8, 2024, doi: 10.1007/s11119-024-10212-2.

[36] Z. Liu *et al.*, “In-season crop yield prediction: State of the art and future research direction,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 146, p. 105129, 2026, doi: https://doi.org/10.1016/j.jag.2026.105129.

[37] J. Zhang, H. Tian, P. Wang, K. Tansey, S. Zhang, and H. Li, “Improving wheat yield estimates using data augmentation models and remotely sensed biophysical indices within deep neural networks in the Guanzhong Plain, PR China,” *Comput. Electron. Agric.*, vol. 192, p. 106616, 2022, doi: https://doi.org/10.1016/j.compag.2021.106616.

[38] I. S. Pothapragada and R. Sujatha, “Smart Agricultural Technology GANs for data augmentation with stacked CNN models and XAI for interpretable maize yield prediction,” *Smart Agric. Technol.*, vol. 11, no. April, p. 100992, 2025, doi: 10.1016/j.atech.2025.100992.

[39] P. Branco, R. P. Ribeiro, L. Torgo, B. Krawczyk, and N. Moniz, “SMOGN: a Pre-processing Approach for Imbalanced Regression,” *Proc. Mach. Learn. Res.*, vol. 74, pp. 36–50, 2017.

ARTICLE IN PRESS

[40] D. Taremwa, E. Ahishakiye, P. K. Kisozi, and F. Kaggwa, “A Weighted Ensemble Learning Approach with Synthetic Oversampling for Maize Yield Modeling in Data-limited Agro-ecological Zones of Uganda,” vol. 10, no. 1, pp. 114–135, 2026.

[41] D. Taremwa, E. Ahishakiye, A. Obbo, P. K. Kisozi, and F. Kaggwa, “Prediction of maize yield in Uganda using CNN-LSTM architecture on a multimodal climate and remote sensing dataset,” *Discov. Artif. Intell.*, 2026, doi: 10.1007/s44163-026-00855-7.

[42] F. Alpsalaz, Y. Özüpak, E. Aslan, and H. Uzel, “Classification of maize leaf diseases with deep learning: Performance evaluation of the proposed model and use of explicable artificial intelligence,” *Chemom. Intell. Lab. Syst.*, vol. 262, p. 105412, 2025, doi: https://doi.org/10.1016/j.chemolab.2025.105412.

[43] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[44] A. Vaswani *et al.*, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[45] S. Kapoor and A. Narayanan, “Leakage and the reproducibility crisis in machine-learning-based science,” *Patterns*, vol. 4, no. 9, 2023, doi: 10.1016/j.patter.2023.100804.

[46] J. Sun, Z. Lai, L. Di, Z. Sun, J. Tao, and Y. Shen, “Multilevel Deep Learning Network for County-Level Corn Yield Estimation in the U.S. Corn Belt,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 5048–5060, 2020, doi: 10.1109/JSTARS.2020.3019046.

[47] U. Bhimavarapu, G. Battineni, and N. Chintalapudi, “Improved Optimization Algorithm in LSTM to Predict Crop Yield,” *Computers*, vol. 12, no. 1, 2023, doi: 10.3390/computers12010010.

[48] M. Qiao, X. He, X. Cheng, P. Li, H. Luo, and L. Zhang, “International Journal of Applied Earth Observations and Geoinformation Crop yield prediction from multi-spectral , multi-temporal remotely sensed imagery using recurrent 3D convolutional neural networks ,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 102, no. April, p. 102436, 2021, doi: 10.1016/j.jag.2021.102436.

[49] M. Shahhosseini, G. Hu, S. Khaki, and S. V. Archontoulis, “Corn Yield Prediction With Ensemble CNN-DNN,” *Front. Plant Sci.*, vol. 12, no. August, pp. 1–13, 2021, doi: 10.3389/fpls.2021.709008.

[50] H. Tian, P. Wang, K. Tansey, D. Han, J. Zhang, and S. Zhang, “International Journal of Applied Earth Observations and Geoinformation A deep learning framework under attention mechanism for wheat yield estimation using remotely sensed indices in the Guanzhong Plain , PR China,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 102, no. 17, p. 102375, 2021, doi: 10.1016/j.jag.2021.102375.

[51] F. Guo *et al.*, “A novel transformer-based neural network under model interpretability for improving wheat yield estimation using remotely sensed multi-variables,” *Comput. Electron. Agric.*, vol. 223, p. 109111, 2024, doi: https://doi.org/10.1016/j.compag.2024.109111.

[52] R. Zeng, J. Li, Z. Li, and Q. Zhang, “Crossformer-Based Model for Predicting and

ARTICLE IN PRESS

Interpreting Crop Yield Variations Under Diverse Climatic and Agricultural Conditions," *Agric.*, vol. 15, no. 9, pp. 1–26, 2025, doi: 10.3390/agriculture15090958.

[53] L. J. Galarza, M. Realpe, M. S. Viñán-ludeña, M. F. Calderón, and S. Jaramillo, "AgriTransformer : A Transformer-Based Model with Attention Mechanisms for Enhanced Multimodal Crop Yield Prediction," pp. 1–16, 2025.

[54] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 3543–3556.

[55] Y. Ma, Z. Zhang, Y. Kang, and M. Özdoğan, *Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach*, vol. 259. 2021. doi: 10.1016/j.rse.2021.112408.

[56] F. Yang *et al.*, "Prediction of corn variety yield with attribute-missing data via graph neural network," *Comput. Electron. Agric.*, vol. 211, p. 108046, 2023, doi: https://doi.org/10.1016/j.compag.2023.108046.

[57] M. Qiao *et al.*, "Spatial-temporal cross-view contrast for self-supervised crop yield prediction," *Expert Syst. Appl.*, vol. 285, p. 128036, 2025, doi: https://doi.org/10.1016/j.eswa.2025.128036.

[58] M. Shahhosseini, G. Hu, I. Huber, and S. V. Archontoulis, "Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt," *Sci. Rep.*, vol. 11, no. 1, pp. 1–15, 2021, doi: 10.1038/s41598-020-80820-1.

[59] B. Cao *et al.*, "Physics-guided deep learning for crop yield estimation," *Eur. J. Agron.*, vol. 172, p. 127850, 2026, doi: https://doi.org/10.1016/j.eja.2025.127850.

[60] J. Lu *et al.*, "Deep Learning for Multi-Source Data-Driven Crop Yield Prediction in Northeast China," *Agric.*, vol. 14, no. 6, 2024, doi: 10.3390/agriculture14060794.

[61] M. Amin *et al.*, "Arti fi cial Intelligence in Agriculture Enhancing crop yield prediction in Senegal using advanced machine learning techniques and synthetic data," *Artif. Intell. Agric.*, vol. 14, pp. 99–114, 2024, doi: 10.1016/j.aiia.2024.11.005.

[62] X. Li *et al.*, "Improving maize yield prediction at the county level from 2002 to 2015 in China using a novel deep learning approach," *Comput. Electron. Agric.*, vol. 202, no. November, p. 107356, 2022, doi: 10.1016/j.compag.2022.107356.

[63] J. Fan, J. Bai, Z. Li, A. Ortiz-Bobea, and C. P. Gomes, "A GNN-RNN Approach for Harnessing Geospatial and Temporal Information: Application to Crop Yield Prediction," *Proc. 36th AAAI Conf. Artif. Intell. AAAI 2022*, vol. 36, pp. 11873–11881, 2022, doi: 10.1609/aaai.v36i11.21444.

[64] F. Lin *et al.*, "MMST-ViT: Climate Change-aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer".

[65] F. Sabo *et al.*, "From Rows to Yields: How Foundation Models for Tabular Data Simplify Crop Yield Prediction," 2025, [Online]. Available: https://arxiv.org/pdf/2506.19046

ARTICLE IN PRESS

[66] X. Li, Y. Lyu, B. Zhu, L. Liu, and K. Song, “Maize yield estimation in Northeast China’s black soil region using a deep learning model with attention mechanism and remote sensing,” *Sci. Rep.*, vol. 15, no. 1, pp. 1–23, 2025, doi: 10.1038/s41598-025-97563-6.

[67] A. M. S. Kheir, A. Govind, V. Nangia, M. Devkota, and A. Elnashar, “Developing automated machine learning approach for fast and robust crop yield prediction using a fusion of remote sensing , soil , and weather dataset Developing automated machine learning approach for fast and robust crop yield prediction using a fusion ,” 2024.

[68] D. Lee *et al.*, “Maize yield forecasts for Sub-Saharan Africa using Earth Observation data and machine learning,” *Glob. Food Sec.*, vol. 33, no. December 2021, p. 100643, 2022, doi: 10.1016/j.gfs.2022.100643.

[69] I. G. Tende, K. Aburada, H. Yamaba, T. Katayama, and N. Okazaki, “Development and Evaluation of a Deep Learning Based System to Predict District-Level Maize Yields in Tanzania,” *Agric.*, vol. 13, no. 3, pp. 1–19, 2023, doi: 10.3390/agriculture13030627.

[70] A. Kaneko *et al.*, “Deep Learning For Crop Yield Prediction in Africa,” *Int. Conf. Mach. Learn. AI Soc. Good Work. Long Beach, United States, 2019*, pp. 1–5, 2019.

[71] D. B. Lobell *et al.*, “Eyes in the Sky, Boots on the Ground: Assessing Satellite- and Ground-Based Approaches to Crop Yield Measurement and Analysis,” *Am. J. Agric. Econ.*, vol. 102, no. 1, pp. 202–219, 2020, doi: 10.1093/ajae/aaz051.

[72] A. Chemura *et al.*, “Contribution of improved varieties to maize productivity under climate change in Uganda,” *Food Secur.*, 2025, doi: 10.1007/s12571-025-01564-2.

[73] J. Bwambale and K. A. Mourad, “Modelling the impact of climate change on maize yield in Victoria Nile Sub-basin, Uganda,” *Arab. J. Geosci.*, vol. 15, no. 1, 2022, doi: 10.1007/s12517-021-09309-z.

[74] A. Agosto, P. Giudici, and A. Tanda, “How to combine ESG scores? A proposal based on credit rating prediction,” *Corp. Soc. Responsib. Environ. Manag.*, vol. 30, no. 6, pp. 3222–3230, Nov. 2023, doi: https://doi.org/10.1002/csr.2548.

[75] M. R. Gupta and A. Sobti, “Time2Agri: Temporal Pretext Tasks for Agricultural Monitoring,” 2025, [Online]. Available: http://arxiv.org/abs/2507.04366

[76] D. Roberts *et al.*, “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure,” *Ecography (Cop.).*, vol. 40, Dec. 2016, doi: 10.1111/ecog.02881.

[77] J. Abate, A. Aman, and D. Adem, “Integration of satellite data for predicting crop yields in Eastern Ethiopia using machine learning,” *Sci. Rep.*, vol. 15, no. 1, pp. 1–12, 2025, doi: 10.1038/s41598-025-00810-z.

[78] J. Pineau *et al.*, “Improving Reproducibility in Machine Learning Research,” *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 1–20, 2021.

[79] D. Paudel *et al.*, “Machine learning for large-scale crop yield forecasting,” *Agric. Syst.*, vol. 187, no. December 2020, p. 103016, 2021, doi: 10.1016/j.agsy.2020.103016.

[80] T. Hu *et al.*, “Crop yield prediction via explainable AI and interpretable machine learning:

ARTICLE IN PRESS

Dangers of black box models for evaluating climate change impacts on crop yield," *Agric. For. Meteorol.*, vol. 336, p. 109458, 2023, doi: https://doi.org/10.1016/j.agrformet.2023.109458.

[81] A. Wolanin *et al.*, "Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt," *Environ. Res. Lett.*, vol. 15, no. 2, 2020, doi: 10.1088/1748-9326/ab68ac.

[82] A. Joshi *et al.*, "An explainable Bi-LSTM model for winter wheat yield prediction," *Front. Plant Sci.*, vol. 15, no. January, pp. 1–17, 2024, doi: 10.3389/fpls.2024.1491493.

[83] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "'Grad-cam: Visual explanations from deep networks via gradient-based localization.' Proceedings of the IEEE international conference on computer vision.," 2017, [Online]. Available: http://gradcam.cloudcv.org

[84] H. Najjar, "Intrinsic explainability of multimodal learning for crop yield simulation," *Comput. Electron. Agric.*, vol. 239, no. PC, p. 111003, 2025, doi: 10.1016/j.compag.2025.111003.

[85] D. Paudel *et al.*, "CY-Bench : A comprehensive benchmark dataset for sub-national crop yield forecasting," no. March, pp. 1–28, 2025.

[86] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *J. Am. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, Mar. 2007, doi: 10.1198/016214506000001437.

[87] P. Giudici and V. Kolesnikov, "SAFE AI metrics: An integrated approach," *Mach. Learn. with Appl.*, vol. 23, p. 100821, 2026, doi: https://doi.org/10.1016/j.mlwa.2025.100821.

[88] M. Shahhosseini, G. Hu, and S. V. Archontoulis, "Forecasting Corn Yield With Machine Learning Ensembles," *Front. Plant Sci.*, vol. 11, no. July, pp. 1–16, 2020, doi: 10.3389/fpls.2020.01120.

[89] M. Mitchell *et al.*, "Model cards for model reporting," *FAT* 2019 - Proc. 2019 Conf. Fairness, Accountability, Transpar.*, no. Figure 2, pp. 220–229, 2019, doi: 10.1145/3287560.3287596.

ARTICLE IN PRESS