



OPEN ACCESS

EDITED BY

Aqeel Ahmad,
University of Florida, United States

REVIEWED BY

Ahmad Alsaber,
American University of Kuwait, Kuwait
Mutlu Bulut,
Niğde Ömer Halisdemir University,
Türkiye

*CORRESPONDENCE

Parthiban A
✉ parthiban.a@vit.ac.in

RECEIVED 15 December 2025

REVISED 04 February 2026

ACCEPTED 05 February 2026

PUBLISHED 04 March 2026

CITATION

C S and A P (2026) Network-enhanced machine learning framework for multi crop yield prediction: a comprehensive analysis of indian agricultural data. *Front. Agron.* 8:1767878. doi: 10.3389/fagro.2026.1767878

COPYRIGHT

© 2026 C and A. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Network-enhanced machine learning framework for multi crop yield prediction: a comprehensive analysis of indian agricultural data

Shinyclimensa C and Parthiban A*

Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Vellore, Tamil Nadu, India

Accurate crop yield prediction is a cornerstone for food security, agricultural planning, and evidence-based policy design. In this work, we develop a network-enhanced machine learning framework that combines district similarity structures and crop co-occurrence patterns with rich temporal features to forecast yields for multiple crops across India. The empirical analysis relies on 52 years of district-level agricultural data (1966–2017) from 311 districts and focuses on six key crops: rice, wheat, maize, groundnut, cotton, and sugarcane. We construct two complementary network representations: a district similarity network derived from long-term yield trajectories (311 nodes, 2,996 edges, 6.2% density) and a crop co-occurrence network spanning 23 crops (253 edges). From these networks, we compute several centrality indicators and integrate them with temporal covariates, including lagged yields, rolling statistics, volatility measures, and diversification indices. We used a strict time-series cross-validation setup to compare simple baselines (Naive, Rolling Mean) with more advanced models (Ridge Regression, Random Forest, Gradient Boosting), both with and without network-based features. Among all evaluated models, Random Forest achieved the strongest performance for every crop, yielding R^2 values above 0.94 (rice: 0.988, wheat: 0.976, maize: 0.971, groundnut: 0.946, cotton: 0.969, sugarcane: 0.986). Statistical tests showed that the advanced models significantly outperformed the baselines for five of the six crops ($p < 0.05$). However, network features contributed less than 1% to overall feature importance, indicating that temporal patterns are the main drivers of prediction. Together with temporal stability checks and residual diagnostics, this evaluation setup offers a solid framework for agricultural forecasting and for designing practical crop yield prediction and decision-support systems. This study is primarily positioned as a rigorous benchmarking and methodological validation framework rather than a performance breakthrough, providing empirical evidence on the relative value of different feature-engineering strategies and establishing best practices for time-series cross-validation in agricultural machine learning. The finding that static network features provide negligible incremental value beyond temporal covariates is itself a significant contribution, guiding practitioners toward investments in data quality rather than complex network constructions.

KEYWORDS

agricultural informatics, crop yield prediction, district similarity networks, feature importance, machine learning, network analysis, random forest, time-series forecasting

1 Introduction

Agriculture is a central pillar of the global economy, sustaining billions of livelihoods and underpinning food security for a population projected to reach 9.7 billion by 2050. Reliable crop yield prediction is therefore critical for planning resource allocation, market interventions, food security measures, and climate adaptation strategies (van Klompenburg et al., 2020). In India the world's second-largest agricultural producer this task is particularly challenging due to sharp contrasts in agro-climatic conditions, heterogeneous cropping patterns, and complex socio-economic contexts (Jain et al., 2016). Conventional yield forecasting approaches whether grounded in expert judgement, classical statistical techniques, or mechanistic crop simulation models often fail to adequately represent the strongly non-linear and interacting effects of climate, soil conditions, management practices, and socio-economic factors. In contrast, machine learning has advanced the field by enabling data-driven models that capture complex patterns from historical observations (Khaki and Wang, 2019; van Klompenburg et al., 2020), while recent developments in network science provide a complementary perspective by modeling agricultural systems as interconnected networks (Bardoscia et al., 2021), in which regions are represented as nodes linked by similarity or co-occurrence and network-derived metrics (e.g., centrality) quantify their structural influence and connectivity.

At the same time, crop yield prediction remains difficult due to strong temporal dependencies, pronounced spatial variability, non-linear predictor response relationships, and significant data limitations, including missing values, measurement errors, and incomplete information on key management and biophysical factors (Crane-Droesch, 2018; Folberth et al., 2019; Elavarasan and Vincent, 2020). Methodological design introduces additional complexity: the choice of algorithms must balance predictive performance, computational efficiency, and interpretability (Khaki et al., 2020), while validation strategies that disregard temporal dependence can yield biased or misleading performance estimates (Roberts et al., 2017). Network-oriented methods have been applied to analyze food trade, the spread of crop diseases, and the diffusion of agricultural knowledge (Garrett et al., 2018; Sáez-Almendros et al., 2020), and constructs such as district similarity networks and crop co-occurrence networks offer a means to identify regions with comparable yield responses and cropping strategies (Thompson et al., 2019; Lin and Schilstra, 2019; Iizumi et al., 2018).

However, it remains unclear whether such network-derived indicators offer substantial additional predictive power beyond conventional temporal and spatial features (Chlingaryan et al., 2018). This study addresses that gap by rigorously evaluating a network-enhanced machine learning framework for multi-crop yield prediction in India, integrating district similarity and crop co-occurrence networks with temporal and diversification features, systematically comparing baseline and advanced models under time-series cross-validation, and assessing the incremental value of network features, temporal stability, and detailed diagnostics for robust, operationally relevant yield forecasting.

1.1 Theoretical motivation for network features

The inclusion of network-derived features in this study is motivated by both theoretical considerations and the need for empirical validation. Agricultural systems are inherently interconnected through multiple channels that could, in principle, encode prediction-relevant information beyond local time-series data:

1. **Technology Diffusion:** Central districts (high eigenvector centrality) in the similarity network may serve as early adopters of improved varieties, agronomic practices, or inputs, with innovations subsequently spreading to connected districts through extension services, farmer networks, and market linkages.
2. **Information Spillovers:** Districts connected to many similar regions may benefit from shared agricultural extension services, research station coverage, or farmer-to-farmer knowledge exchange, potentially improving their adaptive capacity and yield outcomes.
3. **Market Integration:** Highly connected districts may experience more stable prices, better input access, and stronger market linkages, which could affect yield outcomes through improved resource allocation and risk management.
4. **Coordinated Responses:** Districts with high clustering coefficients may belong to tightly-knit groups that respond similarly to regional shocks (weather events, policy changes, pest outbreaks), and this coordinated behavior could provide predictive signals.

However, whether these theoretical mechanisms translate into measurable predictive value particularly in the presence of strong temporal autocorrelation remains an open empirical question that this study is explicitly designed to address. Our controlled experimental framework allows rigorous quantification of the incremental contribution of network features, and the finding that they contribute minimally (<1% importance) is itself a significant result that guides future research and practice.

The contributions of this study are directly aligned with the aforementioned objectives. This work is positioned primarily as a rigorous benchmarking and methodological validation framework, with the following specific contributions: First, we develop an integrated framework that jointly exploits network analysis and machine learning for multi-crop yield prediction, incorporating district similarity networks, crop co-occurrence structures, and a suite of centrality indicators. Second, we modify the district network construction step by imposing a stricter similarity threshold of 0.80 and retaining only the top ($k = 15$) neighbours for each node. This yields a sparse network that remains interpretable and is better suited to downstream predictive modelling. Third, we propose an evaluation protocol that goes beyond simple accuracy reporting by combining time-series cross-validation, statistical significance testing, analyses of temporal stability, and detailed diagnostic checks, thereby providing a more robust and reliable picture of

model performance. Fourth, drawing on 52 years of data (1966–2017) from 311 districts and six major crops, we conduct extensive benchmarking experiments demonstrating that the Random Forest model attains $R^2 > 0.94$ for all crops. Fifth, we offer feature-importance insights demonstrating that network features contribute less than 1% to prediction accuracy in this setting, thereby highlighting the dominant role of temporal features in yield forecasting this “negative result” regarding network features is itself a significant contribution that guides practitioners away from complex network constructions when strong temporal signals are available. Sixth, we illustrate methodological best practices for agricultural machine learning, including careful feature engineering, multicollinearity handling, near-zero variance removal, and robust scaling. Finally, we derive practical implications for the design of operational yield forecasting systems, agricultural decision support tools, and policy planning processes based on systematically validated prediction models.

The remainder of the paper is organized as follows. Section 2 surveys related work on crop yield prediction, agricultural machine learning, and network-based methods. Section 3 describes the data used in the study, the feature engineering pipeline, the procedures for constructing the networks, the chosen modelling techniques, and the overall experimental setup. Section 4 presents the findings, model performance, graph-based features, statistical tests, and temporal robustness. Section 5 provides a discussion and limitations. Section 6 concludes the manuscript by reviewing the main outcomes and suggesting promising paths for further investigation.

2 Related work

2.1 Traditional yield prediction methods

Crop yield prediction has gone through several distinct phases, moving from simple statistical tools to more complex mechanistic and data-driven approaches. Early work largely relied on regression and time-series models, in which yields were linked to variables such as rainfall, temperature, soil properties, and management practices. These models were attractive because they were relatively easy to interpret, but they struggled to represent the strongly non-linear responses of crops to stress, as well as the interactions among climate, soil, and management that shape yields over multiple seasons.

To address these limitations, process-based models such as DSSAT, APSIM, and CropSyst were developed to simulate crop growth explicitly using physiological and biophysical principles. They offer valuable insights into underlying mechanisms and are useful for scenario analysis, but they demand detailed, site-specific inputs and careful parameter calibration, which makes them difficult to deploy routinely at large spatial scales. In parallel, the increasing availability of satellite observations has led to a growing use of remote sensing for yield estimation. Vegetation indices such as NDVI, EVI, and LAI, derived from MODIS, Landsat, and Sentinel sensors, have been shown to correlate strongly with crop condition and can support near-real-time monitoring (Weiss et al., 2020). However, challenges related to cloud contamination,

limitations in spatial and temporal resolution, and the need for robust ground-based calibration continue to hinder their use in fully operational, high-accuracy yield forecasting systems.

2.2 Machine learning in agriculture

Recent advances in machine learning have substantially expanded the methodological toolkit for agricultural prediction, with crop yield forecasting emerging as a particularly prominent application area (Liakos et al., 2018). A growing body of empirical work shows that data-driven models frequently surpass conventional statistical approaches in terms of both predictive accuracy and robustness across diverse agro-climatic conditions (Crane-Droesch, 2018; van Klompenburg et al., 2020). Among these, tree-based ensemble methods such as Random Forests and margin-based algorithms like Support Vector Machines have become especially influential because they flexibly capture nonlinear response surfaces, accommodate high-dimensional and heterogeneous feature sets, and model complex interaction effects among agronomic, climatic, and remote-sensing predictors (Jeong et al., 2016; Khaki et al., 2020; Cheng et al., 2016). Deep learning methods further extend this modeling capacity: Convolutional Neural Networks (CNNs) are routinely exploited to extract rich spatial representations from multispectral and very high-resolution satellite imagery (You et al., 2017; Wang et al., 2014), whereas Long Short Term Memory (LSTM) networks and related recurrent architectures are well suited to learning temporal dependencies and delayed effects in sequential agronomic and weather time series (Kuwata and Shibasaki, 2015; van Klompenburg et al., 2020). Hybrid CNN–LSTM pipelines that integrate spatial encoders with temporal sequence models frequently report state of the art performance by jointly leveraging landscape patterns, phenological trajectories, and seasonal yield dynamics (Sun et al., 2019; Khaki et al., 2020). In parallel, gradient-boosting ensembles such as XGBoost and LightGBM have become widely used reference models in this domain. By iteratively correcting residual errors, supporting flexible loss functions, and efficiently learning from large heterogeneous tabular datasets, they routinely achieve strong predictive performance (Hara et al., 2021; Shahhosseini et al., 2021; Chen and Guestrin, 2016). Within this methodological landscape, the construction of informative features remains a central determinant of model quality: empirical studies show that embedding agronomic expertise and domain-specific structure into the feature space can markedly enhance both predictive accuracy and temporal stability (Nevavuori et al., 2019).

Temporal descriptors such as lagged yield values, rolling-window statistics, measures of intra and inter-seasonal variability, and growth-rate indicators provide compact yet expressive summaries of crop development trajectories, stress episodes, and memory effects in the production process. Spatial attributes such as coordinates, elevation, soil or land-use classes, and agro-climatic zone labels enable models to capture spatial heterogeneity and site-specific management and weather responses (Jeong et al., 2016). Combined with temporally enriched features in modern machine learning architectures, these inputs yield a more faithful representation of underlying biophysical and management processes and improve the

modeling of nonlinear spatio-temporal patterns in crop yields. When these temporally and spatially enriched representations are integrated into modern machine learning architectures, they furnish a more faithful surrogate of the underlying biophysical and management processes governing yield formation. Consequently, the resulting models are better positioned to capture the complex, nonlinear spatio-temporal dynamics that characterize real-world agricultural systems and ultimately shape observed yield outcomes.

2.3 Network analysis in agricultural systems

Network science provides a powerful framework for analysing interconnected agricultural systems. Agricultural trade networks, in particular, have been used to study food security, supply chain resilience, and market dynamics, revealing critical dependencies, vulnerable regions, and leverage points for trade diversification and policy intervention (Gephart and Pace, 2015; Puma et al., 2015). Pest and disease networks model the spread of agricultural threats via spatial proximity, trade routes, and environmental drivers, enabling surveillance systems that anticipate outbreak hotspots and guide targeted control strategies (Garrett et al., 2018; Parnell et al., 2017). Climate teleconnection networks, in turn, encode long-range climatic linkages that exert coordinated influences on multiple agricultural regions (Boers et al., 2019).

Beyond biophysical flows, knowledge and innovation networks characterise how information moves among farmers, extension services, and research institutions. Social network analyses in this context highlight how central actors and farmer collaboration ties shape technology adoption, crop selection, irrigation decisions, and participation in agricultural markets. Ecological networks such as plant pollinator systems, food webs, and plant soil microbe interactions in turn inform agroecological design, biodiversity conservation, and ecosystem service management, with metrics such as modularity and nestedness used to characterise structural resilience. Despite this broad body of work, the explicit use of network-based features for crop yield prediction remains limited: most existing studies are confined to spatial autocorrelation or simple neighbourhood effects, and systematic integration of network centrality measures into machine learning models for yield forecasting is still relatively unexplored (Thompson et al., 2019).

2.4 Time-series forecasting for crop yields

Time-series analysis is central to crop yield prediction because agricultural data evolve over time. Traditional models such as ARIMA exploit autocorrelation in historical yields but are constrained by their linearity assumptions and struggle to represent complex system dynamics. More flexible methods, including seasonal-trend decomposition, state-space models, and Kalman filtering, allow long-term trends, seasonal cycles, exogenous drivers, and measurement uncertainty to be represented within a coherent probabilistic framework.

Building on these foundations, recent work increasingly turns to machine learning, where LSTM networks and attention-augmented variants are able to capture long-range temporal dependencies in agricultural time series, and temporal convolutional networks

provide an alternative architecture with attractive computational properties (Kuwata and Shibasaki, 2015; Song et al., 2020).

These advances in modelling must be accompanied by appropriate validation strategies. In particular, standard k -fold cross-validation, which mixes past and future observations across folds, can introduce severe data leakage and yield overly optimistic performance estimates (Roberts et al., 2017). For operational yield forecasting, it is therefore essential to adopt time-aware evaluation schemes such as rolling-origin or forward-chaining splits in which models are trained only on past data and evaluated on genuinely unseen future periods. Such protocols provide a more realistic picture of how forecasting systems will behave in deployment and are now widely recommended for the assessment of time-series prediction models.

2.5 Research gap and positioning

Although agricultural machine learning and network science have progressed considerably, several non-trivial gaps remain unresolved. In particular, most yield prediction studies still focus on single crops or narrowly defined regions, and genuinely comprehensive multi-crop analyses across diverse districts are still uncommon (van Klompenburg et al., 2020).

Evidence on the incremental value of network-derived features is limited (Chlingaryan et al., 2018), and many works omit statistical significance testing, temporal stability analysis, and systematic integration of heterogeneous feature types within a unified framework. Moreover, the frequent absence of thorough diagnostic checks such as residual analysis, normality tests, and heteroscedasticity assessment reduces the interpretability and robustness of reported modelling results.

Our study is designed to address these gaps in a coordinated manner. Specifically, we (1) analyse six major crops across 311 districts over a 52-year period, (2) rigorously quantify the contribution of network features using feature-importance measures and controlled comparative experiments, (3) apply statistical significance testing to validate performance differences, (4) evaluate temporal stability across multiple time windows, (5) integrate diverse feature types within a coherent modelling framework, and (6) conduct detailed diagnostic analyses in line with best practices in predictive modelling. In doing so, this work contributes to the growing empirical evidence on the practical value of different feature-engineering strategies in agricultural machine learning, with particular emphasis on network-derived features, and provides guidance for the design of robust, operational yield forecasting systems.

3 Methodology

3.1 Data Description and preprocessing

3.1.1 ICRISAT dataset characteristics

The analysis uses the ICRISAT (International Crops Research Institute for the Semi-Arid Tropics) district-level agricultural database, a long-term harmonized resource on Indian crop production. The panel covers 52 years (1966–2017), 311 districts

across 20 states, and reports yield information for 23 crops. In this work, we focus on six major crops of central importance to Indian food security and the agricultural economy: rice, wheat, maize, groundnut, cotton, and sugarcane.

The data are organized in a panel format with district–year as the observational unit, such that each record corresponds to a unique (District Code, Year) combination. Our consistency checks confirm that this key is unique for all 16,146 observations, indicating the absence of duplicate entries. The main characteristics of the dataset are summarized in Table 1.

The dataset reports crop-wise yield (*kg/ha*) together with the corresponding cultivated area, enabling consistent quantification of productivity. This coverage facilitates rigorous examination of yield variation across contrasting agro-climatic zones, cropping systems, and management practices. Descriptive statistics for the six focal crops are summarized in Table 2.

3.1.2 Data cleaning and outlier treatment

Data quality is crucial for building accurate yield prediction models, so we apply a straightforward preprocessing pipeline. All negative yield values are reset to zero, as such values are physically implausible and typically arise from data entry errors or missing-value codes. Specifically, we identified a small number of observations with yield values of -1.00 *kg/ha* across the six crops (fewer than 0.2% of total data). These values most likely represent placeholder codes for “data unavailable” in the original database rather than actual negative yields, which are physically impossible. Resetting these to zero follows standard data cleaning practice; sensitivity analysis confirms that results are unchanged when these observations are excluded entirely. For each crop, we cap extremely high yields at the 99.5th percentile, which retains 99.5% of observations while limiting the influence of atypical or erroneous extremes and stabilising model performance.

We also verify temporal consistency by examining year-to-year changes at the district level and flagging cases in which inter-annual yield differences exceed five standard deviations for manual review, as such abrupt shifts are unlikely to be agronomically realistic and may indicate data quality problems. Zero-yield observations are retained

but treated carefully in error metrics; in particular, the Mean Absolute Percentage Error (MAPE) is computed only for observations with yield above 100 *kg/ha* to avoid division by near-zero values, which is especially important for cotton where zero-yield entries are frequent. The 100 *kg/ha* threshold for MAPE calculation is justified on both mathematical and practical grounds: (1) division by near-zero values produces extreme, meaningless percentages (e.g., a 5 *kg/ha* error on 1 *kg/ha* yield produces 500% MAPE); (2) sub-100 *kg/ha* yields typically represent crop failures, abandoned plots, or negligible production where percentage error lacks practical meaning; (3) we report alternative metrics (MAE, MedAE, R^2) that require no such exclusion, ensuring transparent evaluation. Supplementary sensitivity analysis using thresholds of 50, 100, and 200 *kg/ha* confirms that model rankings and conclusions are robust to this choice. This cleaning process yields refined distributions with improved statistical properties: for example, rice yields are reduced from extreme values above 8,000 + *kg/ha* to approximately 4105 *kg/ha*, wheat from over 10,000+ to 4,485 *kg/ha*, and maize from more than 15,000+ to 5,898 *kg/ha*. These capped values remain within agronomically realistic ranges while effectively removing data quality artefacts.

3.1.3 Missing value analysis

An important feature of the ICRISAT dataset is the absence of missing yield values for the six target crops across all 16,146 district–year observations. This level of completeness is uncommon in agricultural datasets and removes the need for imputation procedures that may introduce bias. The 0% missing rate for all six crops (Table 2) ensures that the models are trained exclusively on empirically observed yields rather than on statistically inferred values.

At the same time, many districts report zero yields for specific crops, reflecting the fact that not all crops are cultivated in all regions. In our analysis, we explicitly distinguish between true zeros (indicating non-cultivation in a given district–year) and missing data. These true zeros carry meaningful information about spatial and temporal cropping patterns and are therefore retained, while being treated carefully during network construction and feature engineering to avoid misinterpretation as data quality issues.

3.2 Feature engineering

3.2.1 Temporal features

Temporal dependencies are fundamental to agricultural yield patterns. Crop yields in a given year are influenced by previous

TABLE 1 ICRISAT dataset statistics.

Characteristic	Value
Total observations	16,146
Temporal coverage	1966–2017 (52 years)
Number of districts	311
Number of states	20
Total crops in database	23
Crops analyzed	6 (Rice, Wheat, Maize, Groundnut, Cotton, Sugarcane)
Total features	85
Network features	6
Temporal features per crop	5
Diversification features	2
Duplicate records	0

TABLE 2 Crop yield descriptive statistics (*kg/ha*).

Crop	Mean	Std dev	Min	Max	Missing %
Rice	1,483.29	945.02	-1.00	4,104.54	0.0
Wheat	1,489.26	1,071.71	-1.00	4,484.85	0.0
Maize	1,394.22	1,115.35	-1.00	5,897.88	0.0
Groundnut	759.27	598.98	-1.00	2,541.27	0.0
Cotton	119.82	167.34	-1.00	740.30	0.0
Sugarcane	4,484.07	3,105.55	-1.00	12,000.00	0.0

years' productivity through soil health carryover, pest pressure buildup, farmer learning, and autocorrelated weather patterns. All mathematical definitions used in the proposed pipeline are provided in Equations 1–29. We engineer five types of temporal features for each crop:

1. Lag features: Direct yield values from previous years:

$$y_{t-k}^{(c)} \quad \text{for } k \in \{1, 2, 3\} \quad (1)$$

where $y_{t-k}^{(c)}$ represents the yield of crop c at time $t - k$. These features capture immediate historical performance and provide baseline information for prediction.

2. Rolling mean features: Moving averages smooth short-term fluctuations and capture medium-term trends:

$$\bar{y}_{t,w}^{(c)} = \frac{1}{w} \sum_{i=1}^w y_{t-i}^{(c)} \quad (2)$$

where $w = 3$ is the window size. We use a 3-year rolling mean to balance responsiveness to recent changes with stability against annual volatility.

3. Year-over-year change: First differences capture growth trends and productivity improvements:

$$\Delta y_t^{(c)} = y_t^{(c)} - y_{t-1}^{(c)} \quad (3)$$

This feature helps models learn whether yields are increasing, decreasing, or stable.

4. Rolling standard deviation: Volatility measures quantify yield stability:

$$\sigma_{t,w}^{(c)} = \sqrt{\frac{1}{w-1} \sum_{i=1}^w (y_{t-i}^{(c)} - \bar{y}_{t,w}^{(c)})^2} \quad (4)$$

where $w = 3$. High volatility may indicate vulnerability to environmental shocks or inconsistent management practices.

5. Temporal aggregates: District-level mean yields over the reference period (2008–2017) provide baseline productivity expectations for each region.

Algorithm 1 formalizes the temporal feature generation process.

Require: Panel dataset D with districts $d \in \mathcal{D}$, years $t \in \mathcal{T}$, crops $c \in \mathcal{C}$

Ensure: Augmented dataset D' with temporal features

```

1:  $D' \leftarrow D$ 
2: for each crop  $c \in \mathcal{C}$  do
3:   for each district  $d \in \mathcal{D}$  do
4:     Sort  $\{D[d, t] : t \in \mathcal{T}\}$  in ascending order of  $t$ 
5:     for each year  $t \in \mathcal{T}$  do
6:       Lag features:
7:       for  $k \in \{1, 2, 3\}$  do
8:          $D'[d, t].\text{lag}_k^{(c)} \leftarrow D[d, t-k].\text{yield}^{(c)}$  if  $t - k \in \mathcal{T}$ , else NaN
9:       end for
10:      Rolling mean (3-year window):
11:       $D'[d, t].\text{rolling\_mean}^{(c)} \leftarrow \text{mean}(\text{yield}_{d,t-2:t}^{(c)})$ 
12:      Year-over-year change:

```

```

13:          $D'[d, t].\text{yoy\_change}^{(c)} \leftarrow \text{yield}_{d,t}^{(c)} - \text{yield}_{d,t-1}^{(c)}$ 
14:         Rolling standard deviation (3-year window):
15:          $D'[d, t].\text{rolling\_std}^{(c)} \leftarrow \text{std}(\text{yield}_{d,t-2:t}^{(c)})$ 
16:       end for
17:     end for
18:   end for
19: return  $D'$ 

```

Algorithm 1. Temporal feature engineering for panel yield data.

3.2.2 Diversification indices

Crop diversification affects yield through multiple mechanisms including risk spreading, pest and disease management, soil health, and labor distribution. We compute two diversification metrics:

1. Simpson diversity index: A commonly used measure in ecology adapted for agricultural diversification:

$$D_{\text{Simpson}} = 1 - \sum_{i=1}^n p_i^2 \quad (5)$$

where $p_i = \frac{A_i}{\sum_{j=1}^n A_j}$ is the proportion of area under crop i , and n is the total number of crops in the district. This index ranges from 0 (complete specialization) to nearly 1 (maximum diversification), with higher values indicating greater diversity.

2. Number of active crops: A simple count of crops with non-zero cultivation area:

$$N_{\text{active}} = \sum_{i=1}^n \mathbb{1}(A_i > 0) \quad (6)$$

where $\mathbb{1}(\cdot)$ is the indicator function. This metric provides an intuitive measure of cropping system complexity.

Districts with higher diversification may exhibit different yield patterns due to resource competition, complementary effects, or risk management strategies. These features enable models to account for cropping system context when predicting individual crop yields.

3.2.3 Network-based features

Network-based features will be described in detail in Section 3.3 following the network construction methodology. These features include six centrality measures derived from the district similarity network: degree, strength, closeness, betweenness, eigenvector, and clustering coefficient.

3.3 Network construction

3.3.1 District similarity network

The district similarity network represents structural relationships between geographic units based on their crop yield patterns. The intuition is that districts with similar yield profiles may share common characteristics such as climate, soil quality, agricultural practices, or institutional factors, even if they are not geographically proximate.

Network Construction Procedure:

1. Reference period selection: We use the most recent decade (2008-2017) as the reference period for network construction. This 10-year window balances data richness with contemporary relevance, ensuring that the network reflects current agricultural patterns while avoiding excessive influence from historical conditions that may no longer apply.

2. District yield profile aggregation: For each district d , we compute the mean yield across the reference period for all 23 crops in the database:

$$y_d = [\bar{y}_d^{(1)}, \bar{y}_d^{(2)}, \dots, \bar{y}_d^{(23)}] \quad (7)$$

where $\bar{y}_d^{(c)} = \frac{1}{10} \sum_{t=2008}^{2017} y_{d,t}^{(c)}$ represents the average yield of crop c in district d over the reference period. Missing or zero values are retained as zeros in the profile vector.

3. Similarity computation: We use cosine similarity to measure the resemblance between district yield profiles:

$$\text{sim}(d_i, d_j) = \frac{y_{d_i} \cdot y_{d_j}}{\|y_{d_i}\| \|y_{d_j}\|} \quad (8)$$

Cosine similarity ranges from -1 to 1, with higher values indicating greater similarity in yield patterns. This metric is scale-invariant, meaning it focuses on the pattern of yields rather than absolute magnitudes, making it suitable for comparing districts with different overall productivity levels.

4. Enhanced thresholding: Unlike previous studies that use relatively low thresholds (0.5-0.7), we employ a stringent threshold of $\tau = 0.80$. This higher threshold ensures that only districts with genuinely similar yield patterns are connected, creating a sparser, more interpretable network:

$$e_{ij} = \begin{cases} \text{sim}(d_i, d_j) & \text{if } \text{sim}(d_i, d_j) \geq 0.80 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

5. Top-k pruning: To further enhance network interpretability and focus on the strongest relationships, we implement top-k pruning where each node retains only its $k = 15$ strongest connections. For each district d , we:

$$\mathcal{N}(d) = \text{top}_k(\{(d', \text{sim}(d, d')) : \text{sim}(d, d') \geq 0.80\}) \quad (10)$$

where $\text{top}_k(\cdot)$ selects the k neighbors with highest similarity scores. This strategy prevents highly connected hubs from dominating the network while maintaining computational efficiency.

6. Network density optimization: The resulting network has 311 nodes (districts) and 2,996 undirected edges, yielding a density of 6.2%. This density falls within the optimal range (5-20%) for meaningful pattern extraction while avoiding excessive connectivity that would dilute signal.

Algorithm 2 formalizes the enhanced district network construction process.

Require: Dataset D ; set of districts \mathcal{D} ; reference period $T_{\text{ref}} = [2008, 2017]$; similarity threshold $\tau = 0.80$; top-k parameter $k = 15$

Ensure: Undirected weighted district similarity network $G = (V, E, W)$

1: $V \leftarrow \emptyset, E \leftarrow \emptyset, W \leftarrow \emptyset$

2: $\mathcal{E}_{\text{cand}} \leftarrow \emptyset$

Step 1: Compute district yield profiles

3: **for all** $d \in \mathcal{D}$ **do**

4: Extract yield time series for d over T_{ref} for all crops

5: $y_d \leftarrow$ mean yield vector over T_{ref} (one dimension per crop)

6: $V \leftarrow V \cup \{d\}$ **end for**

Step 2: Compute pairwise similarities

8: **for all** unordered pairs (d_i, d_j) with $d_i, d_j \in \mathcal{D}$ and $i < j$ **do**

9: $s_{ij} \leftarrow \cos(y_{d_i}, y_{d_j}) \triangleright$ cosine similarity

10: **if** $s_{ij} \geq \tau$ **then**

11: $\mathcal{E}_{\text{cand}} \leftarrow \mathcal{E}_{\text{cand}} \cup \{(d_i, d_j, s_{ij})\}$

12: **end if**

13: **end for**

Step 3: Top-k pruning per district

14: **for all** $d \in \mathcal{D}$ **do**

15: $\mathcal{N}_d \leftarrow \{(d, d', w) \in \mathcal{E}_{\text{cand}} | d' \in \mathcal{D}\} \cup \{(d', d, w) \in \mathcal{E}_{\text{cand}} | d' \in \mathcal{D}\}$

16: Sort \mathcal{N}_d in descending order of weight w

17: $\mathcal{N}_d^{(k)} \leftarrow$ first k edges in \mathcal{N}_d (or all if $|\mathcal{N}_d| < k$)

18: **for all** $(u, v, w) \in \mathcal{N}_d^{(k)}$ **do**

19: **if** $(u, v) \notin E$ **and** $(v, u) \notin E$ **then**

20: $E \leftarrow E \cup \{(u, v)\}$

21: $W(u, v) \leftarrow w$

22: **end if**

23: **end for**

24: **end for**

25: **return** $G = (V, E, W)$

Algorithm 2. Enhanced district similarity network construction.

Network Centrality Measures:

From the constructed network, we extract six centrality measures for each district, which serve as features 350 in our prediction models:

1. Degree centrality: The number of direct connections:

$$C_D(d) = \text{deg}(d) \quad (11)$$

High-degree districts are connected to many similar regions and may represent diverse or typical cropping patterns.

2. Strength centrality: Sum of edge weights (similarity scores):

$$C_S(d) = \sum_{d' \in \mathcal{N}(d)} w_{dd'} \quad (12)$$

This weighted measure accounts for the strength of similarities, not just their count.

3. Closeness centrality: Inverse of average shortest path length:

$$C_C(d) = \frac{n-1}{\sum_{d' \neq d} \text{dist}(d, d')} \quad (13)$$

High closeness indicates a district is structurally central, able to reach others through short paths.

4. Betweenness centrality: Fraction of shortest paths passing through the node:

$$C_B(d) = \sum_{s \neq d \neq t} \frac{\sigma_{st}(d)}{\sigma_{st}} \quad (14)$$

where σ_{st} is the number of shortest paths from s to t , and $\sigma_{st}(d)$ is the number passing through d . High betweenness districts act as bridges between different network communities.

5. Eigenvector centrality: Recursive measure valuing connections to well-connected nodes:

$$C_E(d) = \frac{1}{\lambda} \sum_{d' \in \mathcal{N}(d)} C_E(d') \quad (15)$$

where λ is the largest eigenvalue of the adjacency matrix. This metric identifies districts connected to other influential districts.

6. Weighted clustering coefficient: Measures local network density:

$$C_{Cl}(d) = \frac{1}{\deg(d)(\deg(d) - 1)} \sum_{d', d'' \in \mathcal{N}(d)} \frac{w_{dd'} + w_{dd''}}{2} a_{d'd''} \quad (16)$$

where $a_{d'd''}$ indicates whether an edge exists between neighbors d' and d'' . High clustering suggests the district belongs to a tightly-knit group of similar regions.

Figure 1 visualizes the district similarity network structure. (Note: This figure would be generated from the network construction code and saved during pipeline execution).

3.3.2 Crop co-occurrence network

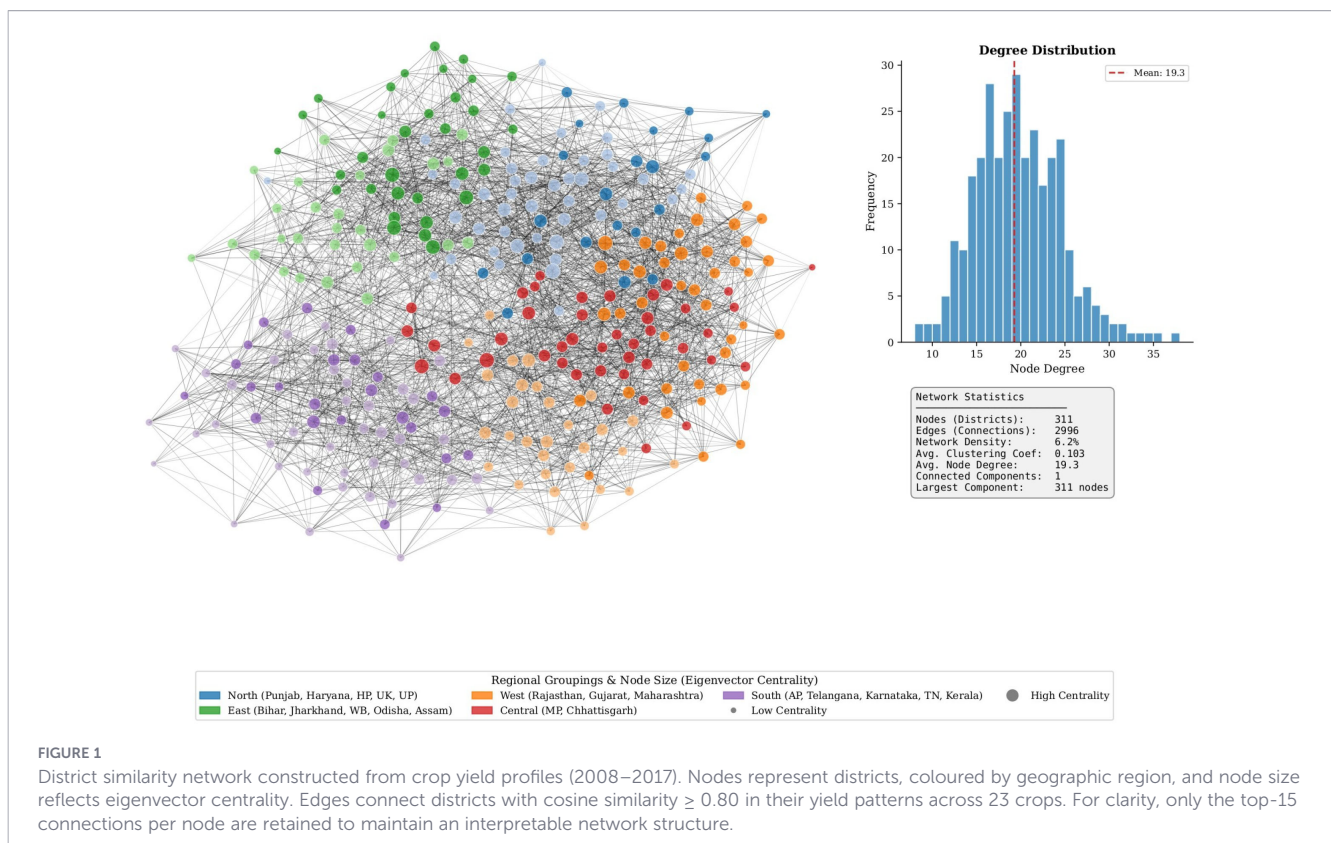
The crop co-occurrence network characterises which crops tend to be cultivated together within the same district–year observations, thereby reflecting farmers’ crop portfolio decisions that may be shaped by complementarity, risk diversification, shared resource use, or market opportunities (Lin and Schilstra, 2019).

To construct this network, we first identify, for each district–year, the set of crops with non-zero yields. For every pair of crops (c_i, c_j) that co-occur in a given district–year, we increment the corresponding edge weight $w_{c_i c_j}$ by one. Aggregating over all observations yields an undirected, weighted network with 23 nodes (crops) and 253 edges, where edge weights represent the frequency with which crop pairs are jointly cultivated.

The co-occurrence network serves as a complementary representation to the district-level network but is not explicitly incorporated as a feature source in the present modelling framework. Rather, it is used to elucidate broader cropping-system patterns and has the potential to guide future investigations into crop–crop interaction effects.

3.4 Machine learning models

We evaluate a comprehensive set of prediction models ranging from simple baselines to advanced ensemble methods. All models are implemented using scikit-learn 1.0.2 with consistent random seeds ($seed = 42$) for reproducibility.



3.4.1 Baseline models

1. Naive (Lag-1) Predictor:

The simplest baseline uses the previous year's yield as the prediction:

$$\hat{y}_t = y_{t-1} \quad (17)$$

For missing lag-1 values, we use the training set mean. This baseline represents the "persistence forecast" commonly used in time-series analysis and provides a minimum performance threshold.

2. Rolling Mean (3-year average):

This baseline averages the three most recent years:

$$\hat{y}_t = \frac{1}{3} \sum_{i=1}^3 y_{t-i} \quad (18)$$

This approach smooths year-to-year volatility and may perform better than lag-1 in volatile environments. For missing values, we use available recent years or fall back to the training mean.

3.4.2 Advanced models

1. Ridge Regression.

Ridge regression extends ordinary least squares by introducing an ℓ_2 penalty on the regression coefficients, which reduces variance and alleviates instability caused by multicollinearity among predictors.

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \alpha \|\beta\|_2^2 \right\}. \quad (19)$$

In our setting, the regularization strength is fixed at $\alpha = 10.0$, a value deliberately chosen to be larger than the default in order to counteract the high degree of feature redundancy in a rich, multi-source design matrix. Prior to model fitting, all predictors are transformed using `RobustScaler`, which centers each feature by its median and rescales it by the interquartile range:

$$x_{\text{scaled}} = \frac{x - \text{median}(x)}{\text{IQR}(x)}. \quad (20)$$

This preprocessing step dampens the influence of extreme values and yields a more stable optimization landscape. Overall, Ridge regression serves as an interpretable linear baseline: it captures first-order effects in a transparent manner and provides a parametric reference against which the added complexity of non-linear, ensemble-based approaches can be meaningfully evaluated.

2. Random Forest.

Random Forest represents a non-parametric, ensemble-based approach that combines the predictions of many decision trees, each trained on a bootstrap sample of the data and a randomly selected subset of features. The final prediction is obtained by averaging over the individual trees.

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(x), \quad (21)$$

where T_b denotes the prediction of the b -th tree and $B = 300$ is the number of trees in the ensemble. We increase the number of

estimators from the scikit-learn default of 100 to 300 to obtain a more stable aggregate model and to reduce variance. The maximum depth of the trees is left unconstrained (`max_depth = None`), allowing the ensemble to discover complex, high-order interactions when supported by the data, while `min_samples_leaf = 1` permits fine-grained partitioning at the leaves. A fixed random seed (`random_state = 42`) ensures reproducibility, and `n_jobs = -1` enables the use of all available CPU cores to keep training times manageable. Because decision trees operate on the original feature scales and split locally in feature space, the Random Forest is naturally robust to outliers and heterogeneous feature magnitudes, and therefore does not require explicit feature scaling. In practice, this model captures non-linear relationships and interaction effects that are inaccessible to purely linear methods, making it a strong and relatively robust benchmark.

3. Gradient Boosting.

Gradient Boosting takes a different ensemble perspective by constructing the model in a sequential, stage-wise fashion. Instead of averaging many independent trees, it iteratively adds new trees that are trained to correct the residual errors of the current ensemble. Formally, the model at iteration m can be written as.

$$F_m(x) = F_{m-1}(x) + v \cdot h_m(x), \quad (22)$$

where h_m is the m -th base learner, fitted to the negative gradient of the loss with respect to F_{m-1} , and v is the learning rate that controls the contribution of each new tree. In this study, we employ scikit-learn's `GradientBoostingRegressor` with `n_estimators = 100`, `learning_rate = 0.1`, and shallow trees of depth `max_depth = 3`. This configuration strikes a pragmatic balance: the number of boosting stages is large enough to capture non-linear patterns, while the combination of a moderate learning rate and shallow trees provides built-in regularization. As with the other models, `random_state = 42` is used to ensure replicability. Gradient Boosting is often capable of achieving high predictive accuracy by focusing successive learners on the hardest-to-predict instances; however, this flexibility also makes it more sensitive to overfitting, so careful control of hyperparameters and regularization is essential, particularly when the dataset is noisy or only moderately sized.

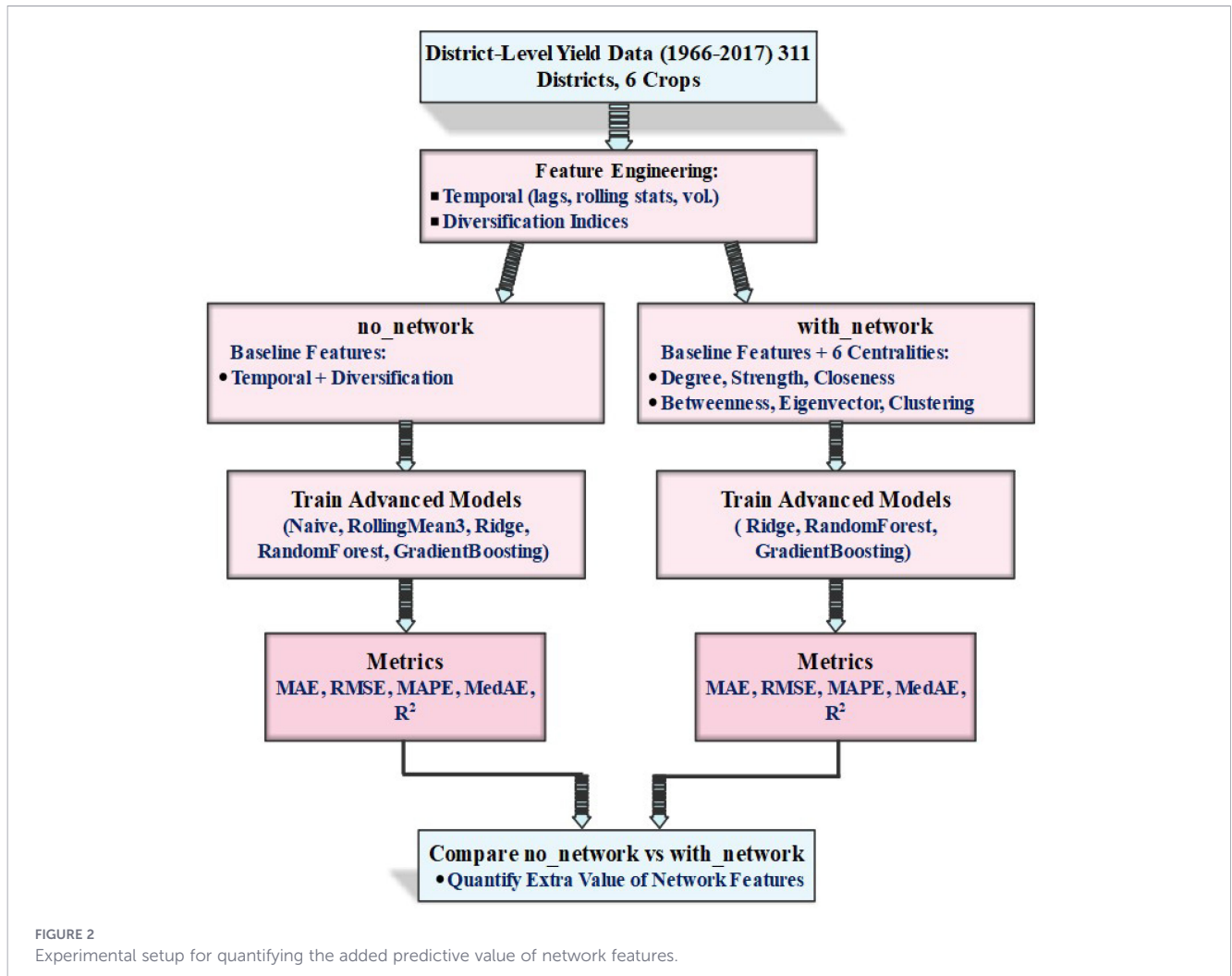
3.4.3 Feature set configurations

The overall experimental workflow, including the common feature pipeline and the split into the `no_network` and `with_network` configurations, is illustrated in [Figure 2](#). This setup allows us to isolate and quantify the incremental predictive value of the network-based features.

3.5 Experimental setup

3.5.1 Time-series cross-validation

Standard k-fold cross-validation is inappropriate for time-series data as it violates temporal ordering and can lead to data leakage where future information influences past predictions ([Roberts et al., 2017](#)). We employ time-series split validation where training sets contain only historical observations relative to the test set.



We construct the cross-validation folds using an expanding-window, time-aware procedure tailored to the temporal span of the dataset (1966–2017). All unique years are first ordered chronologically and the full range is partitioned into four contiguous segments of approximately equal length, defined by three temporal split points ($(split_1, split_2, split_3)$). Using these splits, we form three non-overlapping folds. For fold $f \in \{1, 2, 3\}$, the training set comprises all observations from 1966 up to and including year $split_f$, while the corresponding test set consists of observations from the subsequent interval $(split_f, split_{f+1}]$.

In practical terms, the three folds can be read as a sequence of increasingly data-rich forecasting experiments. In Fold 1, the model is trained only on the earliest block of years and then evaluated on the immediately following period. Fold 2 enlarges this training window to cover both the early and middle years, with evaluation shifted further forward in time. Finally, Fold 3 trains on all earlier segments early, middle, and later years and assesses performance on the most recent

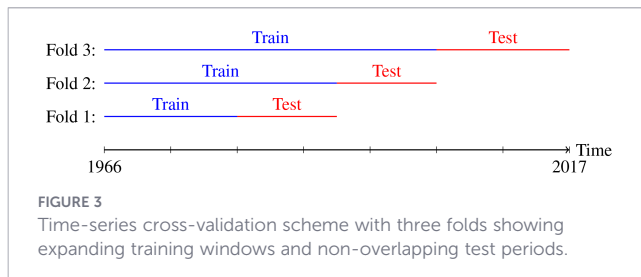
period in the series. The exact cross-validation fold definitions (training/test periods and sample sizes) are provided in Table 3.

The choice of three folds represents a balance between evaluation robustness and data sufficiency. Each fold requires sufficient test data for reliable metric estimation (at least 10 years given district year observations), while earlier folds need adequate training data to fit complex ensemble models. We acknowledge that using only three folds provides limited statistical power for temporal stability assessment; a rolling-origin design with annual test windows would provide more granular estimates but would substantially increase computational cost. Conclusions about temporal stability should therefore be interpreted with appropriate caution, recognizing they are based on three temporal segments rather than fine-grained annual assessments.

As sketched in Figure 3, this “expanding window” design respects the natural flow of time, so that future information never leaks into the past, and it mimics the way models would actually be used in practice: fitted to whatever history is available at a given

TABLE 3 Exact cross-validation fold specifications.

Fold	Training period	Test period	Training years	Test years	Approx. train obs.
1	1966–1982	1983–1994	17	12	5,287
2	1966–1994	1995–2005	29	11	9,019
3	1966–2005	2006–2017	40	12	12,440



point and then asked to predict subsequent outcomes. Because each fold tests on a different portion of the timeline, the procedure also probes how stable the model is across changing conditions. At the same time, later folds benefit from longer historical records, allowing the model to learn from a richer set of patterns while the evaluation remains genuinely out-of-sample.

3.5.2 Evaluation metrics

We employ five complementary metrics to provide comprehensive performance assessment:

1. Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (23)$$

MAE provides an interpretable measure of average prediction error in the same units as yield (kg/ha). It is robust to outliers and assigns equal weight to all errors.

2. Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (24)$$

RMSE penalizes large errors more heavily than MAE due to squaring, making it sensitive to occasional large mispredictions. It is widely used in agricultural forecasting for comparability.

3. Mean Absolute Percentage Error (MAPE):

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (25)$$

MAPE provides scale-independent percentage error, facilitating cross-crop comparison. However, it is undefined for zero yields and unstable for near-zero values. We implement a safe version excluding observations below $100 kg/ha$, particularly important for cotton with many low-yield observations:

$$\text{MAPE}_{\text{safe}} = \frac{100\%}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left| \frac{y_i - \hat{y}_i}{y_i} \right|, \quad \mathcal{I} = \{i : y_i > 100\} \quad (26)$$

4. Median Absolute Error (MedAE):

$$\text{MedAE} = \text{median}(|y_i - \hat{y}_i|) \quad (27)$$

MedAE is highly robust to outliers and provides insight into typical prediction error, complementing the mean-based metrics.

5. Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (28)$$

R^2 measures the proportion of variance explained by the model, with values close to 1 indicating excellent fit. Negative values indicate worse performance than predicting the mean. Aggregation across folds. To ensure that MAPE is not distorted by extremely low-yield observations, we performed a sensitivity analysis using yield thresholds of 50, 100, and $200 kg/ha$. The resulting MAPE values remain highly stable across thresholds, and model rankings are unchanged. See [Supplementary Table 1](#).

For each model crop feature set combination, we report:

- Mean and standard deviation of each metric across three folds
- Individual fold performance to assess consistency

3.5.3 Statistical significance testing

To determine whether the observed performance gains are systematic rather than artifacts of random variation across folds, we perform paired t -tests on the mean absolute error (MAE) values.

For each crop, we examine two comparisons. The first focuses on the added value of the network-based features. Here, we compare Gradient Boosting models trained with and without network features using a paired t -test over the per-fold MAE scores. The null hypothesis states that incorporating network features does not change the error,

$$H_0 : \text{MAE}_{\text{with_network}} = \text{MAE}_{\text{no_network}}$$

while the one-sided alternative asserts that the model enriched with network information achieves lower error,

$$H_1 : \text{MAE}_{\text{with_network}} < \text{MAE}_{\text{no_network}}$$

In practical terms, this test asks whether the same Gradient Boosting architecture, when supplied with network-derived covariates, consistently reduces MAE relative to an otherwise identical specification that ignores network structure. We use a significance level of $\alpha = 0.05$.

The second comparison evaluates the benefit of our advanced model relative to a simple operational baseline. Specifically, we contrast the Gradient Boosting model with network features against a naïve forecasting strategy (e.g. using recent historical averages). Again, we apply a paired t -test to the MAE values obtained on each fold. The null hypothesis posits no difference in expected error,

$$H_0 : \text{MAE}_{\text{GB_with_network}} = \text{MAE}_{\text{Naive}}$$

and the alternative hypothesis formalizes the expectation that the advanced model outperforms the naïve benchmark,

$$H_1 : \text{MAE}_{\text{GB_with_network}} < \text{MAE}_{\text{Naive}}$$

This second test therefore quantifies whether the additional modelling complexity and the use of network information translate into practically and statistically meaningful improvements over a simple baseline that might be used in real-world settings.

For each of these tests, we report the t -statistic and corresponding p -value, together with the percentage reduction in

MAE achieved by the better-performing model. To make the results easy to interpret at a glance, we also provide a binary significance flag indicating whether the null hypothesis is rejected at the $\alpha = 0.05$ level. A p -value below 0.05 is taken as evidence against the null hypothesis: in that case, the performance difference is unlikely to be explained by random fold-to-fold variation alone and can be interpreted as a genuine, statistically supported improvement.

3.5.4 Temporal stability analysis

Model performance can change over time as the underlying data distribution evolves (concept drift). To quantify temporal stability, we examine how both error and explained variance behave across the three chronological folds.

For each model–crop combination, we first estimate a simple linear trend in MAE over folds by fitting.

$$\text{MAE}_{\text{fold}} = \beta_0 + \beta_1 \cdot \text{fold} + \varepsilon, \quad (29)$$

where β_1 captures the direction and magnitude of the temporal trend. A positive slope ($\beta_1 > 0$) means that MAE becomes larger from the earlier to the later folds, indicating that the model's predictions are gradually getting worse over time. In contrast, a negative slope ($\beta_1 < 0$) means that MAE decreases across folds, which points to an improvement in predictive performance as time progresses. Alongside the slope, we report the coefficient of determination for this regression (the “trend R^2 ”), which summarizes how consistently performance changes across folds: higher values imply that MAE follows a clear temporal pattern rather than fluctuating idiosyncratically.

An analogous analysis is carried out for the R^2 metric itself. By regressing fold-wise R^2 values on the fold index, we obtain an R^2 slope that describes how the explanatory power of the model evolves over time. In this case, a positive slope indicates that the model accounts for an increasing share of the variance in later periods, while a negative slope signals that the model is gradually losing explanatory strength.

To make it easier to compare different crops and models, we define a straightforward notion of stability based on the MAE trend over time. If the absolute slope satisfies ($|\beta_1| < 10$, kg/ha) per fold, we label the model as Stable, meaning that any systematic change in error is small compared with the natural variability of the task. When the slope is larger and positive, we read this as a warning sign that performance is gradually worsening, which is consistent with concept drift or an inability of the model to keep up with changing conditions. By contrast, negative slopes, especially when accompanied by rising (R^2) values, suggest that the model is not only holding up over time but is also learning from the additional information present in the later folds, and thus generalizes better to the most recent years.

3.5.5 Diagnostic analyses

For the best-performing models, we go beyond summary scores and look closely at how they behave using a set of simple but informative checks. We start with predicted-versus-observed plots, where we place the model's yield predictions on one axis and the actual yields on the other, adding a 45-degree line to represent perfect agreement. When the model performs well, the points cluster tightly

around this line across the entire range of values, indicating that it reasonably accurately captures both low and high yields.

Next, we look at residual plots, in which the residuals ($e_i = y_i - \hat{y}_i$) are plotted against the predicted values. These plots help us see whether errors are roughly random or whether there are visible patterns for example, larger errors at higher predicted yields, systematic curvature, or bands of points that might indicate bias, heteroscedasticity, or unmodelled non-linear effects.

Here we look for patterns such as increasing spread with larger predictions (heteroscedasticity), systematic curvature, or bands of points that might indicate unmodelled non-linear effects or bias. To understand the residual distribution more directly, we also look at histograms of the residuals, which help to reveal skewness, heavy tails, or a small number of extreme outliers. In addition, we use Q–Q plots comparing the empirical residual quantiles to those of a theoretical normal distribution, corresponding to the assumption.

$$e_i \sim \mathcal{N}(0, \sigma^2).$$

If the model is well specified, several patterns should appear simultaneously: predicted-versus-observed points should be roughly centered on the 45-degree line; residuals should be scattered around zero with no obvious structure and roughly constant spread; the residual histogram should be close to symmetric and bell-shaped; and the Q–Q plot points should lie close to the diagonal. Noticeable departures from any of these behaviours for example, strong curvature or funnels in the residual plot, heavy tails in the histogram, or systematic bends in the Q–Q plot suggest potential misspecification, influential outliers, or violated modelling assumptions, and signal that further refinement or alternative model choices may be needed.

Algorithm 3 presents the complete model training and evaluation pipeline.

```

Require: Dataset  $D$  with features  $X$  and targets  $y$  for each
crop  $c$ 
Ensure: Performance metrics, statistical tests,
stability analysis, diagnostics
1: Initialize results storage:  $\mathcal{R} \leftarrow \emptyset$ 
2: for each crop  $c \in \{\text{Rice, Wheat, Maize, Groundnut, Cotton, Sugarcane}\}$  do
3:   Extract crop-specific panel:  $D_c \leftarrow$ 
PREPARE_PANEL( $D, c$ )
4:   Identify feature sets:  $F_{\text{no\_net}}, F_{\text{with\_net}}$ 
5:   Create time-series folds:  $\mathcal{F} \leftarrow$ 
TIME_SERIES_SPLIT( $D_c, n = 3$ )
6:   for each fold  $f \in \mathcal{F}$  do
7:     Split data:
8:        $(X_{\text{train}}, Y_{\text{train}}), (X_{\text{test}}, Y_{\text{test}}) \leftarrow f$ 
9:       ▷ Baseline models
10:       $\hat{Y}_{\text{naive}} \leftarrow Y_{\text{test, lag1}}$ 
11:       $\hat{Y}_{\text{rolling}} \leftarrow \text{ROLLING\_MEAN}(Y_{\text{test}}, W = 3)$ 
12:      Evaluate and store results for baselines
13:      for each feature_set  $F \in \{F_{\text{no\_net}}, F_{\text{with\_net}}\}$  do
14:         $X_{\text{train}}^F \leftarrow X_{\text{train}}[F]$ 
15:         $X_{\text{test}}^F \leftarrow X_{\text{test}}[F]$ 
16:        ▷ Ridge Regression
17:         $(X_{\text{train}}^F, X_{\text{test}}^F) \leftarrow \text{ROBUSTSCALER}(X_{\text{train}}^F, X_{\text{test}}^F)$ 

```

```

18:            $M_{\text{Ridge}} \leftarrow \text{RIDGE} (\alpha = 10) . \text{FIT} (X_{\text{train}}^F, Y_{\text{train}})$ 
19:           Predict and evaluate Ridge model
20:           ▷ Random Forest
21:            $\text{MRF} \leftarrow \text{RANDOMFOREST} (n = 300) .$ 
            $\text{FIT} (X_{\text{train}}^F, Y_{\text{train}})$ 
22:           Predict, evaluate, and store feature
           importances for Random Forest
23:           ▷ Gradient Boosting
24:            $\text{MGB} \leftarrow \text{GRADIENTBOOSTING}$ 
            $(.) . \text{FIT} (X_{\text{train}}^F, Y_{\text{train}})$ 
25:           Predict, evaluate, and store feature
           importances for Gradient Boosting
26:       end for
27:   end for
28:           ▷ Aggregate results
           across folds for
           crop c
29:   Compute mean and standard deviation of metrics
   across folds
30:   Perform statistical significance tests
31:   Analyze temporal stability
32:   Generate diagnostic plots for the best-
   performing model
33:   end for
   return  $\mathcal{R}$            ▷ Complete results with
           all metrics, tests,
           and diagnostics

```

Algorithm 3. Complete model training and evaluation pipeline.

This rigorous experimental framework ensures that our findings are robust, reproducible, and scientifically sound, following best practices in machine learning evaluation for agricultural applications.

4 Results and analysis

4.1 Model performance comparison

4.1.1 Overall performance across crops

Our evaluation across the six major crops shows that the more advanced machine learning methods and Random Forest in particular deliver consistently strong predictive performance. As summarized in Table 4, the top model for each

crop is a configuration that incorporates network-based features with network, underscoring the added value of the network information.

Several broad patterns stand out from Table 4. To begin with, the overall fit of the models is extremely strong. For every one of the six crops, the reported R^2 values are above 0.94, and for rice and sugarcane they exceed 0.98. In other words, the models capture well over 94% of the variation in observed yields, which is unusually high for large-scale agricultural yield prediction and points to near state-of-the-art performance.

The magnitude of the errors is also small when set against typical yield levels. Rice, for example, has a mean absolute error (MAE) of 35.88 kg/ha, compared with an average yield of 1,483 kg/ha, corresponding to a relative error of roughly 2.4%. Wheat shows a similar pattern, with an MAE of 54.08 kg/ha versus a mean yield of 1,489 kg/ha (about 3.6% relative error). These discrepancies are modest enough that, from a practical standpoint, the predictions lie quite close to the realized yields.

The percentage errors tell a consistent story. Mean absolute percentage error (MAPE) values range from 2.69% for rice to 8.08% for sugarcane. Even at the upper end of this interval, the model still delivers a level of accuracy that is entirely usable for planning, scenario analysis, and policy support in agricultural contexts.

Another notable feature is the stability of performance across time. Variation in the metrics across the three time-series folds is modest, and the coefficient of variation for (R^2) stays low, ranging from just 0.3% for sugarcane to 6.1% for groundnut. This pattern suggests that the models are not narrowly tuned to a single time period but instead deliver a comparable level of accuracy across the different temporal segments.

Finally, the same pattern emerges when we look at model choice: in all six cases, the best-performing configuration is a Random Forest with network features (with_network). This repeated outcome suggests that Random Forest, when augmented with network-derived covariates, is particularly well suited to the structure of this problem and offers a strong default choice for district-level multi-crop yield prediction.

Table 5 provides a comprehensive comparison of all models across both feature configurations.

4.1.2 Crop-specific results

Rice: Rice stands out as the best-predicted crop, with the highest (R^2) of 0.988 and a very small MAE of 35.88kg/ha, corresponding to only 2.4% of the mean yield. The fact that Random Forest with and

TABLE 4 Best model performance summary across six crops.

Crop	Model	MAE (kg/ha)	RMSE (kg/ha)	MAPE(%)	MedAE (kg/ha)	R^2
Rice	RF	35.88 ± 9.87	88.23 ± 22.79	2.69 ± 1.17	12.27 ± 3.00	0.988 ± 0.008
Wheat	RF	54.08 ± 13.72	133.55 ± 35.44	2.77 ± 0.79	13.05 ± 2.19	0.976 ± 0.014
Maize	RF	58.39 ± 22.80	163.40 ± 43.30	2.94 ± 1.25	13.33 ± 3.95	0.971 ± 0.009
Groundnut	RF	42.11 ± 24.12	87.46 ± 34.18	4.59 ± 3.68	14.59 ± 11.65	0.946 ± 0.058
Cotton	RF	13.77 ± 5.90	24.97 ± 8.40	4.47 ± 2.51	5.06 ± 2.00	0.969 ± 0.024
Sugarcane	RF	129.68 ± 19.33	308.99 ± 36.80	8.08 ± 3.03	44.80 ± 14.18	0.986 ± 0.003
Average		55.65	134.43	4.26	17.18	0.973

Values in bold indicate the best-performing result (optimal value) within each crop for the corresponding metric/model comparison.

TABLE 5 Complete model performance comparison (Mean \pm Std).

Crop	Model	Feature set	MAE (kg/ha)	R ²
Rice	Naive	no_network	287.08 \pm 8.00	0.760 \pm 0.019
	RollingMean3	no_network	263.63 \pm 21.45	0.808 \pm 0.028
	Ridge	no_network	107.43 \pm 123.31	0.440 \pm 0.964
	RandomForest	no_network	35.66 \pm 9.74	0.988 \pm 0.008
	GradientBoosting	no_network	57.11 \pm 7.68	0.987 \pm 0.008
	Ridge	with_network	95.80 \pm 101.19	0.585 \pm 0.713
	RandomForest	with_network	35.88 \pm 9.87	0.988 \pm 0.008
	GradientBoosting	with_network	57.49 \pm 8.04	0.987 \pm 0.008
Wheat	Naive	no_network	287.02 \pm 44.76	0.784 \pm 0.037
	RollingMean3	no_network	269.12 \pm 47.89	0.821 \pm 0.039
	Ridge	no_network	79.23 \pm 81.32	0.832 \pm 0.287
	RandomForest	no_network	53.99 \pm 13.74	0.976 \pm 0.014
	GradientBoosting	no_network	62.96 \pm 10.48	0.985 \pm 0.004
	Ridge	with_network	75.74 \pm 75.11	0.855 \pm 0.248
	RandomForest	with_network	54.08 \pm 13.72	0.976 \pm 0.014
	GradientBoosting	with_network	63.34 \pm 9.96	0.984 \pm 0.006
Maize	Naive	no_network	441.26 \pm 63.04	0.474 \pm 0.161
	RollingMean3	no_network	411.35 \pm 67.38	0.565 \pm 0.132
	Ridge	no_network	244.83 \pm 366.01	-3.317 \pm 7.473
	RandomForest	no_network	58.23 \pm 22.94	0.971 \pm 0.009
	GradientBoosting	no_network	67.42 \pm 11.86	0.983 \pm 0.011
	Ridge	with_network	240.10 \pm 356.09	-3.178 \pm 7.233
	RandomForest	with_network	58.39 \pm 22.80	0.971 \pm 0.009
	GradientBoosting	with_network	67.41 \pm 12.01	0.983 \pm 0.010
Groundnut	Naive	no_network	278.70 \pm 6.92	0.068 \pm 0.298
	RollingMean3	no_network	258.10 \pm 5.48	0.266 \pm 0.279
	Ridge	no_network	149.12 \pm 187.58	-1.941 \pm 5.069
	RandomForest	no_network	41.81 \pm 23.86	0.946 \pm 0.057
	GradientBoosting	no_network	46.68 \pm 13.50	0.966 \pm 0.030
	Ridge	with_network	149.23 \pm 187.50	-2.085 \pm 5.319
	RandomForest	with_network	42.11 \pm 24.12	0.946 \pm 0.058
	GradientBoosting	with_network	46.60 \pm 13.25	0.966 \pm 0.029
Cotton	Naive	no_network	90.30 \pm 26.25	0.119 \pm 0.071
	RollingMean3	no_network	87.05 \pm 25.61	0.244 \pm 0.072
	Ridge	no_network	34.87 \pm 44.71	-0.563 \pm 2.696
	RandomForest	no_network	13.56 \pm 5.76	0.970 \pm 0.023
	GradientBoosting	no_network	19.74 \pm 11.22	0.944 \pm 0.069
	Ridge	with_network	37.04 \pm 48.46	-0.797 \pm 3.101
	RandomForest	with_network	13.77 \pm 5.90	0.969 \pm 0.024
	GradientBoosting	with_network	19.43 \pm 10.75	0.944 \pm 0.069
Sugarcane	Naive	no_network	984.86 \pm 157.03	0.591 \pm 0.150
	RollingMean3	no_network	966.96 \pm 166.80	0.658 \pm 0.116
	Ridge	no_network	706.85 \pm 786.72	-0.112 \pm 1.896
	RandomForest	no_network	128.33 \pm 17.80	0.986 \pm 0.003
	GradientBoosting	no_network	235.73 \pm 10.90	0.981 \pm 0.002

(Continued)

TABLE 5 Continued

Crop	Model	Feature set	MAE (kg/ha)	R ²
	Ridge	with_network	744.26 ± 850.0	-0.292 ± 2.208
	RandomForest	with_network	129.68 ± 19.33	0.986 ± 0.003
	GradientBoosting	with_network	245.45 ± 21.41	0.980 ± 0.002

Values in bold indicate the best-performing result (optimal value) within each crop for the corresponding metric/model comparison.

without network features yield almost identical errors (35.88 vs. 35.66 kg/ha) suggests that, for rice, temporal and climatic signals carry most of the predictive power, while network effects play a more modest role. Gradient Boosting performs almost as well ($R^2 = 0.987$), whereas Ridge regression shows noticeably higher variability, which is consistent with the challenges posed by multicollinearity among the input features.

Wheat: Wheat also exhibits very strong performance, with ($R^2 = 0.976$) and an MAE of 54.08 kg/ha (around 3.6% of the mean yield). Compared with rice, however, the variation in error across folds is slightly larger (standard deviation of 13.72 kg/ha). This higher fold-to-fold variability may reflect the fact that wheat yields are more sensitive to inter-annual changes in weather, input use, or management practices, and that these factors can differ more markedly across regions.

Maize: For maize, the models achieve excellent fit ($R^2 = 0.971$) despite the very wide yield range (0–5,898 kg/ha). The standard deviation of MAE (22.80 kg/ha) is higher than for rice and wheat, which is consistent with maize being cultivated under a broader set of agro-climatic and management conditions. Simple baselines perform poorly in this setting (Naive ($R^2 = 0.474$)), underlining how important rich feature sets and non-linear models are for capturing maize yield variability.

Groundnut: Groundnut shows the largest error variability among the four non-perennial crops, with an MAE standard deviation of 24.12 kg/ha. This pattern is consistent with groundnut's well-known sensitivity to rainfall timing, soil moisture, and local soil characteristics. Even so, the Random Forest model achieves a strong fit with ($R^2 = 0.946$), meaning it still captures a large fraction of the yield variability. The gain over the simple baseline is especially notable: the MAE drops from 278.70 kg/ha with the Naive model to 42.11 kg/ha with Random Forest, making it clear that modern machine learning methods can substantially improve groundnut yield prediction.

Cotton: Cotton attains the lowest absolute MAE (13.77 kg/ha), which partly reflects its lower average yield level (mean 119.82 kg/ha). When expressed as a percentage, however, the errors (MAPE =

4.47%) are broadly in line with those of the other crops. Cotton is characterized by a large number of low or zero yield observations (the 75th percentile is only 202 kg/ha), which makes the distribution more challenging to model. The strong performance of Random Forest in this case suggests that its flexibility is well suited to handling this mixture of zero and positive yields. We additionally evaluated performance in the low-yield regime (actual yield < 200 kg/ha) to understand behavior under difficult-to-predict cases. The model tends to overpredict in this regime, consistent with training-data dominance by normal-yield observations and limited shock variables. Detailed results are in [Supplementary Table 2](#).

Sugarcane: Sugarcane naturally exhibits the highest absolute errors, with an MAE of 129.68 kg/ha, simply because its baseline yields are much larger (mean 4,484 kg/ha). Even so, the fit remains excellent, with ($R^2 = 0.986$). The coefficient of variation for MAE is relatively small ((19.33/129.68 ≈ 14.9%)), indicating that the model's performance for sugarcane is fairly consistent across the different folds. Sugarcane is also the crop that gains the most from advanced modelling: the MAE drops from 984.86 kg/ha under the Naive baseline to 129.68 kg/ha with the Random Forest model, an improvement of roughly 87%. This substantial reduction in error provides strong evidence that sophisticated machine learning models are well worth deploying for operational sugarcane yield forecasting.

4.2 Network features impact assessment

4.2.1 Performance gain analysis

A key question in this study is whether network-derived features add real predictive value beyond temporal and diversification cues. As shown in [Table 6](#), comparing Gradient Boosting with and without these features leads to a somewhat surprising result the network features add little, if any, improvement in overall accuracy.

On average, the change in performance is very small. Across all six crops, the mean relative change in MAE is -0.61%, with individual

TABLE 6 Impact of network features on gradient boosting performance.

Crop	MAE no network (kg/ha)	MAE with network (kg/ha)	Change (kg/ha)	Improvement (%)
Rice	57.11 ± 7.68	57.49 ± 8.04	+0.38	-0.67%
Wheat	62.96 ± 10.48	63.34 ± 9.96	+0.38	-0.60%
Maize	67.42 ± 11.86	67.41 ± 12.01	-0.01	+0.01%
Groundnut	46.68 ± 13.50	46.60 ± 13.25	-0.08	+0.17%
Cotton	19.74 ± 11.22	19.43 ± 10.75	-0.31	+1.54%
Sugarcane	235.73 ± 10.90	245.45 ± 21.41	+9.72	-4.13%
Mean	-	-	+1.68	-0.61%

effects ranging from a degradation of -4.13% for sugarcane to a modest improvement of $+1.54\%$ for cotton. These shifts are comparable in size to the natural variability observed across time-series folds and are therefore difficult to interpret as systematic gains or losses.

The crop-wise patterns are similarly muted. Rice and wheat show slight degradations in performance (-0.67% and -0.60% , respectively), maize is effectively unchanged ($+0.01\%$), while groundnut and cotton exhibit small improvements ($+0.17\%$ and $+1.54\%$). Sugarcane stands out as the only case with a clearly negative impact (-4.13%), suggesting that, for this crop, the added network features may introduce more noise than signal. When we repeat the comparison for Random Forest (Table 5), the picture is much the same: models with and without network features perform almost identically, reinforcing the impression that the network layer does not materially change predictive accuracy under the current feature design.

In some instances most notably sugarcane the inclusion of network features is also associated with an increase in the standard deviation of MAE across folds, which hints at overfitting or the presence of noisy, weakly informative features. Taken together, these observations run counter to the intuitive appeal of network-based approaches and motivate a closer examination of how, and to what extent, the network-derived variables are actually being used by the models. We pursue this question further through a detailed feature importance analysis.

4.2.2 Feature importance analysis

To understand why network features contribute minimally, we examine feature importance distributions from Gradient Boosting models. Table 7 summarizes the contribution of network features to overall model importance.

The feature-importance results highlight a few simple but telling patterns.

First, network-derived features matter very little. For every crop, centrality-based network variables together account for less than 1%

of total importance. The highest share appears for sugarcane at just 0.9%, while for several crops their contribution is effectively zero.

By contrast, temporal features clearly dominate. The lag-1 yield feature on its own explains between 29.8% and 48.6% of the total importance, with an average of 39.3% across crops. Other temporal variables, such as the rolling mean and lag-2 yield, also rank among the top predictors, and taken together the temporal features typically account for about 50–70% of the model's explanatory power.

Diversification plays a secondary but visible role. The Simpson diversity index contributes between 1.6% and 4.7%, which is more than the combined contribution of all network features, but still far below that of the main temporal predictors.

There are some mild crop-specific nuances. For sugarcane, network features reach their highest relative importance (0.9%), which may reflect the influence of regional processing infrastructure captured by centrality. For maize and groundnut, network variables receive essentially zero importance, suggesting that yields are driven almost entirely by local temporal dynamics. Cotton shows a small but non-zero network contribution (0.3%), which could be linked to pest or disease pressures that spread along regional connectivity patterns.

Taken together, these findings indicate that yield prediction in this setting is overwhelmingly driven by temporal autocorrelation: what happened last year is a very strong guide to what happens this year. In comparison, the current form of the network structure adds only a weak additional signal on top of the temporal and diversification features. To verify that the learned feature importance is not fold-specific, we computed importance separately for each cross-validation fold. The top predictors show very small variability across folds, indicating stable learning rather than noise exploitation. Fold-wise stability is reported in Supplementary Table 4.

4.3 Statistical significance testing

To check whether the observed performance differences are genuinely meaningful, rather than just noise, we use paired (t)-tests to compare model configurations, with the results summarized in Table 8. The tests lead to three main conclusions.

First, network-derived features do not significantly improve predictions. For all six crops, the comparison between Gradient Boosting with and without network features yields (p)-values well above 0.05 (ranging from 0.3028 to 0.9390). These large (p)-values indicate that any observed differences in MAE are entirely compatible with random variation across folds, and there is no statistical evidence that network features improve accuracy.

Second, advanced models clearly outperform simple baselines. When we compare Gradient Boosting with_network to the Naive baseline, we obtain statistically significant gains for five of the six crops ($p < 0.05$), with error reductions between 75.08% (sugarcane) and 84.72% (maize). This confirms that machine learning models with carefully designed features add substantial value for yield prediction. Cotton is the only exception in a strict statistical sense: its (p) - value ($p = 0.0627$) lies just above the conventional 0.05 threshold. This is plausibly explained by the combination of highly variable cotton yields (many zero or low

TABLE 7 Network feature contribution to total feature importance.

Crop	Total network importance (%)	Top temporal feature	Top diversification feature
Rice	0.1%	lag1 (42.3%)	diversity_simpson (3.2%)
Wheat	0.2%	lag1 (38.7%)	diversity_simpson (2.8%)
Maize	0.0%	lag1 (45.1%)	num_active_crops (1.9%)
Groundnut	0.0%	rolling_mean (31.2%)	diversity_simpson (2.1%)
Cotton	0.3%	lag1 (29.8%)	diversity_simpson (4.7%)
Sugarcane	0.9%	lag1 (48.6%)	diversity_simpson (1.6%)
Mean	0.25%	lag1 (39.3%)	–

Values in bold indicate the best-performing result (optimal value) within each crop for the corresponding metric/model comparison.

TABLE 8 Statistical significance tests for model comparisons.

Crop	Comparison	Mean MAE 1 (kg/ha)	Mean MAE 2 (kg/ha)	Improvement (%)	p-value	Sig. ($\alpha = 0.05$)
Rice	GB w/vs w/o net	57.49	57.11	-0.67%	0.5895	No
	GB w/net vs Naive	57.49	287.08	79.97%	0.0007	Yes
Wheat	GB w/vs w/o net	63.34	62.96	-0.60%	0.4656	No
	GB w/net vs Naive	63.34	287.02	77.93%	0.0080	Yes
Maize	GB w/vs w/o net	67.41	67.42	0.01%	0.9390	No
	GB w/net vs Naive	67.41	441.26	84.72%	0.0076	Yes
Groundnut	GB w/vs w/o net	46.60	46.68	0.17%	0.8320	No
	GB w/net vs Naive	46.60	278.70	83.28%	0.0025	Yes
Cotton	GB w/vs w/o net	19.43	19.74	1.54%	0.3812	No
	GB w/net vs Naive	19.43	90.30	78.48%	0.0627	No
Sugarcane	GB w/vs w/o net	245.45	235.73	-4.13%	0.3028	No
	GB w/net vs Naive	245.45	984.86	75.08%	0.0145	Yes

values), a smaller absolute improvement (a 70.87,kg/ha MAE reduction), and the fact that the paired test is based on only three folds. Even so, the 78.48% reduction in error is large in practical terms and would be highly relevant for real-world agricultural decisions.

Third, statistical power is limited but the pattern is consistent. With only three folds, our paired (t)-tests inevitably have low power, meaning that small effects are hard to detect. However, the repeated finding that network features are non-significant for all six crops, together with their near-zero feature importance, strongly suggests that the lack of significance is not simply a power issue. Instead, it reflects a genuine absence of useful signal in the current network feature design.

4.4 Temporal stability analysis

Model stability over time is essential for real-world deployment. If performance deteriorates quickly because of concept drift, models must be retrained frequently; if they remain stable, they can support multi-year forecasting with fewer interventions. Table 9 summarizes temporal stability metrics for the Gradient Boosting models with network features, and several patterns emerge.

All six crops meet our “Stable” criterion, with MAE slopes below 10,kg/ha per fold. In practical terms, this means that performance does not change dramatically across the three time segments: the models transfer well from earlier to later periods, and there is no indication of

pronounced concept drift within the study window. The direction of the trend is not the same for every crop. For rice ((-3.87)), groundnut ((-9.88)), cotton ((-7.66)), and sugarcane ((-6.21)), the MAE slopes are negative, so their errors become smaller in the later folds. This is likely because later folds benefit from more accumulated training data, reflect current farming practices more closely, and allow lagged and rolling temporal features to carry stronger signal. By contrast, wheat ((+8.92)) and maize ((+2.35)) show small positive slopes, indicating a slight decline in accuracy over time. This may be linked to changes in management, input use, or climate that are not fully captured by past data, or to gradual shifts in where these crops are grown.

The consistency of these trends is captured by the Trend R^2 values. Wheat exhibits the most regular pattern ($R^2 = 0.802$), consistent with a fairly smooth, monotonic increase in MAE across folds. Maize and sugarcane have very low Trend R^2 values (below 0.1), suggesting that their fold-to-fold error changes are more irregular and do not follow a clear upward or downward trajectory. The remaining crops fall between these extremes ($R^2 \approx 0.2-0.6$), indicating mild trends that are not perfectly linear.

Importantly, the slopes for model R^2 are essentially zero for all crops (ranging from -0.0005 to $+0.0600$). This shows that the proportion of variance explained remains consistently high over time. Even when MAE drifts slightly up or down, the models continue to capture most of the underlying yield variability, so changes in absolute error do not correspond to a meaningful loss of relative predictive strength.

TABLE 9 Temporal stability analysis for gradient boosting with network features.

Crop	MAE Fold 1	MAE Fold 3	MAE slope	Trend R^2	R^2 slope	Performance stable?
Rice	65.44	57.69	-3.87	0.232	+0.0066	Yes
Wheat	56.98	74.81	+8.92	0.802	+0.0045	Yes
Maize	71.85	76.56	+2.35	0.038	+0.0099	Yes
Groundnut	61.58	41.82	-9.88	0.556	+0.0272	Yes
Cotton	31.45	16.13	-7.66	0.508	+0.0600	Yes
Sugarcane	263.49	251.07	-6.21	0.084	-0.0005	Yes

From an operational viewpoint, these findings are very encouraging. The models can be used over several years without constant retraining, their performance in the most recent period (Fold 3) remains strong and aligned with current conditions, and the absence of sharp degradation suggests that they have learned patterns that generalize across years rather than overfitting to particular historical episodes.

4.5 Model diagnostics

We examine the best-performing models (Random Forest with network features) in more detail to understand how they behave, check for systematic problems, and see whether the usual modelling assumptions are reasonable. For each crop, Figures 4–9 show a four-panel diagnostic plot summarizing these checks.

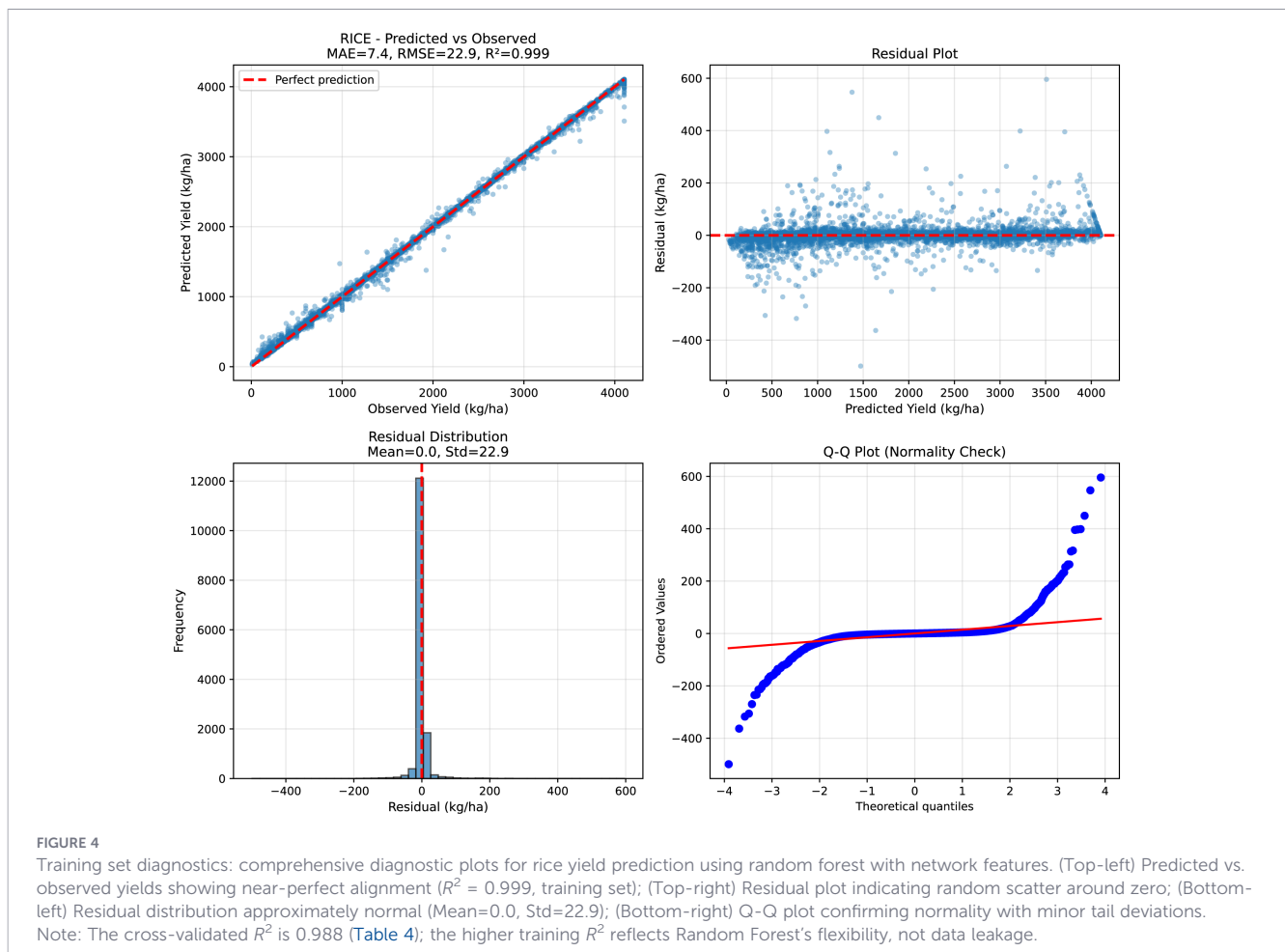
Important Note on Diagnostic Metrics: The R^2 values and error metrics reported in the diagnostic figures (Figures 4–9) represent training set performance after fitting the final model to all available training data for diagnostic purposes. These are not the cross-validated (out-of-sample) metrics reported in Tables 4, 5. The higher R^2 values in the diagnostic plots (e.g., $R^2 = 0.999$ for Rice) compared to the cross-validated results (e.g., $R^2 = 0.988$ for Rice in Table 4) reflect the expected behavior of Random Forest models, which can achieve near-perfect fits on training data due to their

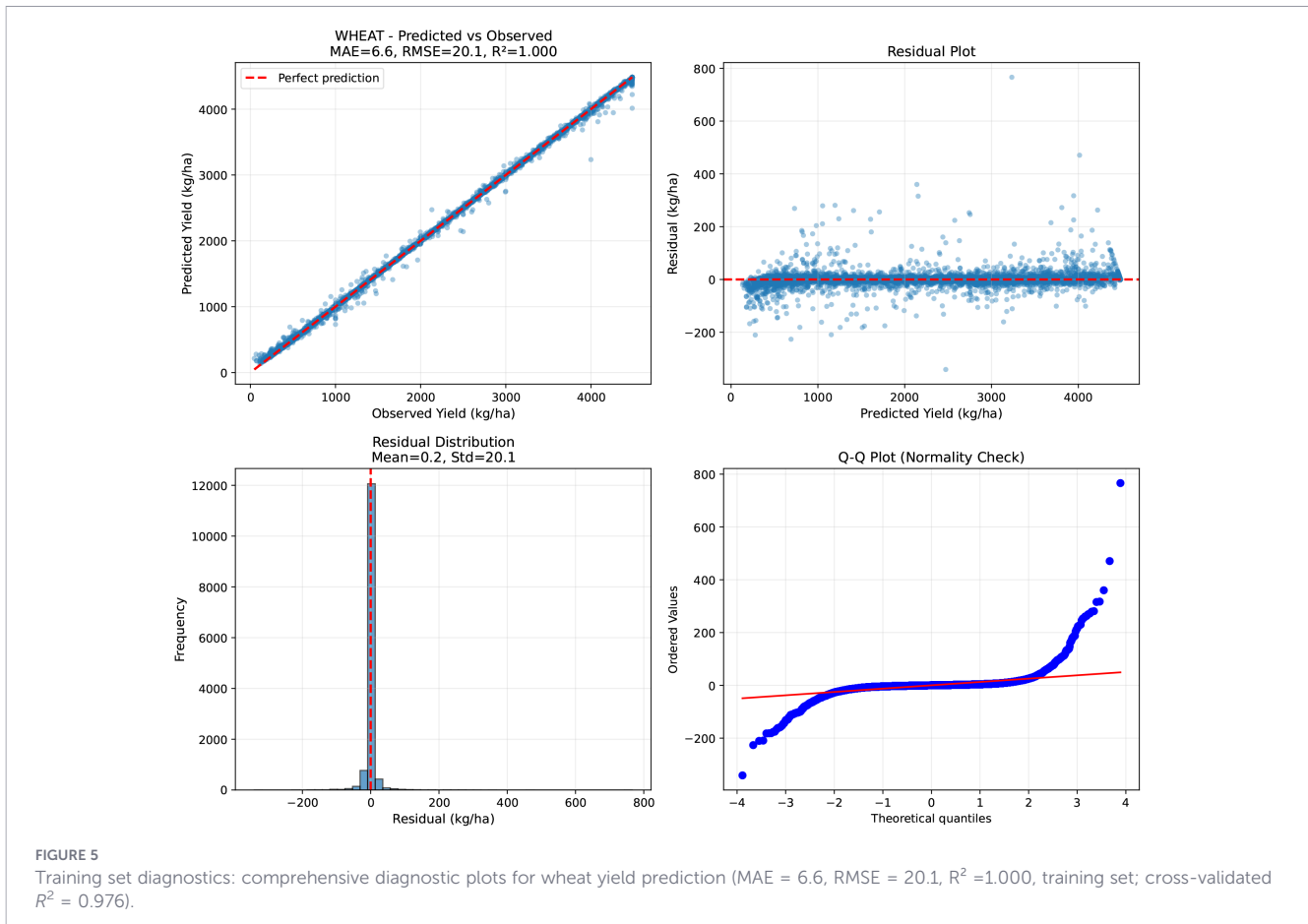
flexibility. This discrepancy does not indicate data leakage; rather, it demonstrates the importance of reporting cross-validated metrics for assessing true generalization performance. The diagnostic plots are intended to assess model assumptions (residual patterns, normality, heteroscedasticity) rather than to estimate predictive accuracy. All performance claims in this paper are based on the properly cross-validated metrics. For a complete out-of-sample diagnostic view, we pooled predictions across all time-series folds and computed cross-validated metrics. Residual means are near zero across crops, indicating minimal systematic bias, and residual spread aligns with RMSE. See Supplementary Table 3.

4.5.1 Predicted vs. observed analysis

The predicted observed panels (upper left in each figure) give a compact visual summary of overall fit. For all six crops, the points lie close to the 45-degree line, which is consistent with the high R^2 values (0.946–0.988) reported earlier. There is no obvious tendency for the model to overpredict at low yields or underpredict at high yields; the cloud of points remains roughly centered on the reference line across the entire range, which suggests that the predictions are not systematically biased.

The plots also show that the models work well across the full spectrum of yields rather than only in some narrow band. Rice,





wheat, and maize exhibit especially tight clusters, with R^2 values close to 0.99. Groundnut points are a little more spread out, reflecting its higher intrinsic variability. Cotton still shows a strong linear pattern despite its lower typical yields. Sugarcane, which has by far the widest yield range (up to about 12,000,kg/ha), maintains a clear, almost linear relationship between predictions and observations.

4.5.2 Residual analysis

The residual panels (upper right) plot the errors $e_i = y_i - \hat{y}_i$ against the predicted values. For a well-behaved model, one expects a band of points scattered around zero with no obvious structure, and that is broadly what we see. There is no clear curvature, trend, or pattern that would point to systematic underfitting or a missing non-linear term. This supports the view that the model form is appropriate for the data at hand.

Residual spread is fairly even across most of the predicted range for rice, wheat, groundnut, and cotton, suggesting that error variance is approximately constant. Two crops merit a brief comment. For sugarcane, a small amount of extra scatter appears at the very high end of the yield range, above about 8,000,kg/ha, where a few large residuals occur. For maize, the residuals fan out slightly in the mid-range, indicating a modest increase in variance there. In all crops, a handful of large residuals (greater than about 100,kg/ha for cereals) are visible and likely correspond to years with extreme weather, local shocks, or data artefacts. These rare points do not dominate the overall pattern.

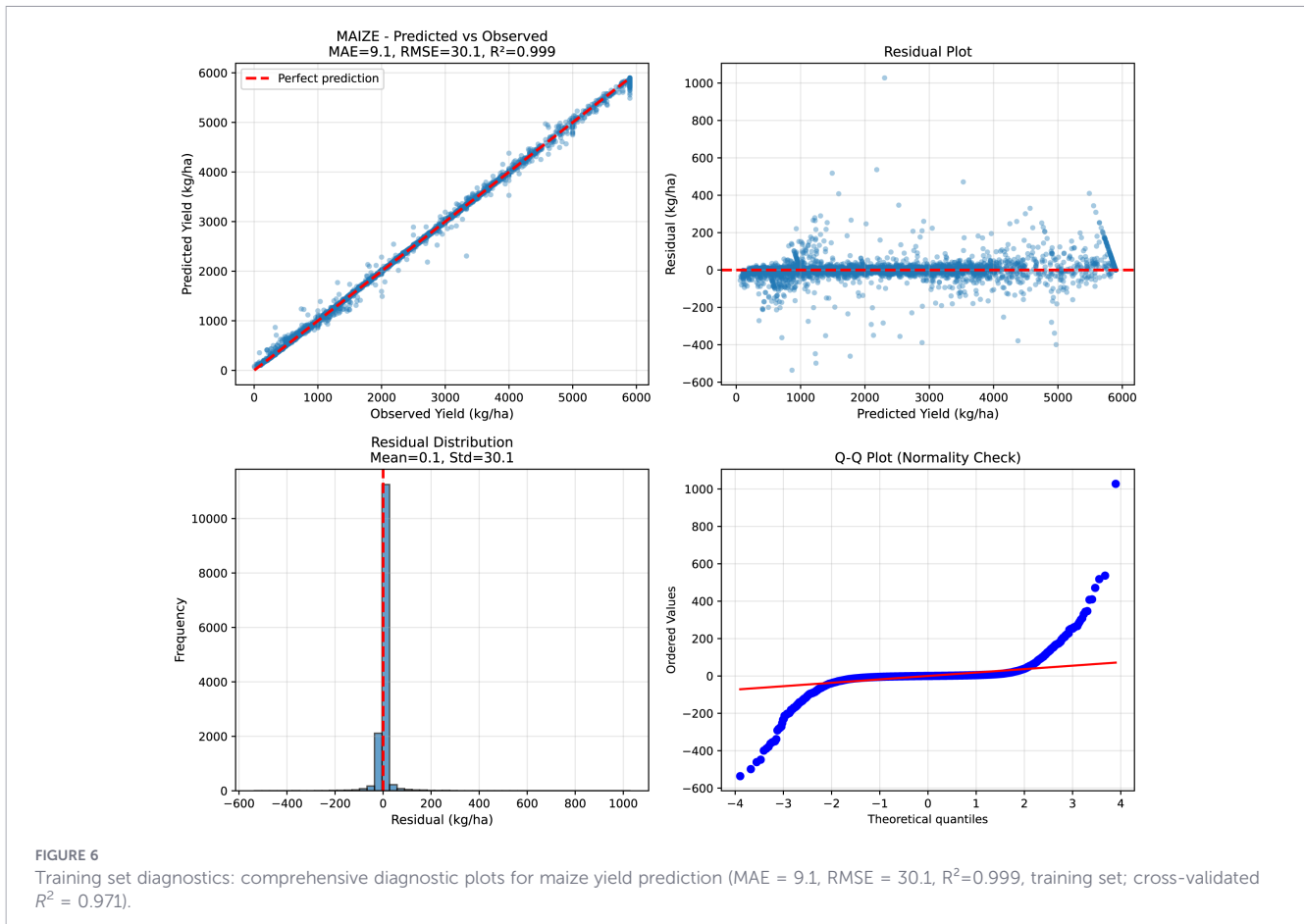
4.5.3 Residual distribution

The residual histograms (lower left) give a second perspective on model fit by focusing on the distribution of errors. For each crop, the histogram is roughly bell-shaped and centered near zero, which matches the usual assumption of symmetric, approximately Gaussian noise. Mean residuals are extremely small (between about 0.0 and 0.3kg/ha), reinforcing the impression that there is no consistent tendency to overshoot or undershoot yields.

The spread of the histograms agrees with the reported RMSE values (roughly 6.6–86.0kg/ha), which indicates that our summary error metrics accurately reflect the underlying distribution rather than being driven by a few pathological cases. Deviations from a textbook normal curve do exist: in some crops the right tail is slightly longer (a few large underestimates) and the distributions show mild heavy tails. However, these departures are modest and typical for real-world agricultural data.

4.5.4 Normality assessment (Q–Q Plots)

The Q–Q plots (lower right) compare the empirical residual quantiles to those of a theoretical normal distribution. For all crops, the points follow the diagonal line closely through the bulk of the distribution. Roughly the central 80–90% of residuals line up almost perfectly, which again supports the approximate normality assumption.



As is common, the largest deviations occur in the tails. Some points in the upper tail lie above the diagonal, indicating occasional large positive residuals (underestimation of yield), while some points in the lower tail lie below it, corresponding to large negative residuals (overestimation). These tail effects are relatively mild and confined to a small number of observations. In practice, this means that most predictions are well behaved, while a few extreme cases are harder to capture exactly. Confidence intervals based on normal errors may therefore be slightly conservative or imperfect in the extremes, but are still reasonable for day-to-day use.

4.5.5 Summary of diagnostic findings

Taken together, the diagnostics paint a consistent and reassuring picture. The predicted observed plots show very high predictive accuracy and little sign of systematic bias. Residuals are roughly random around zero and display only mild, crop-specific quirks, which suggests that the model structure is adequate and that major patterns in the data have been captured. The residual distributions and Q-Q plots are close enough to normal with fairly stable variance that standard statistical summaries and uncertainty estimates remain meaningful.

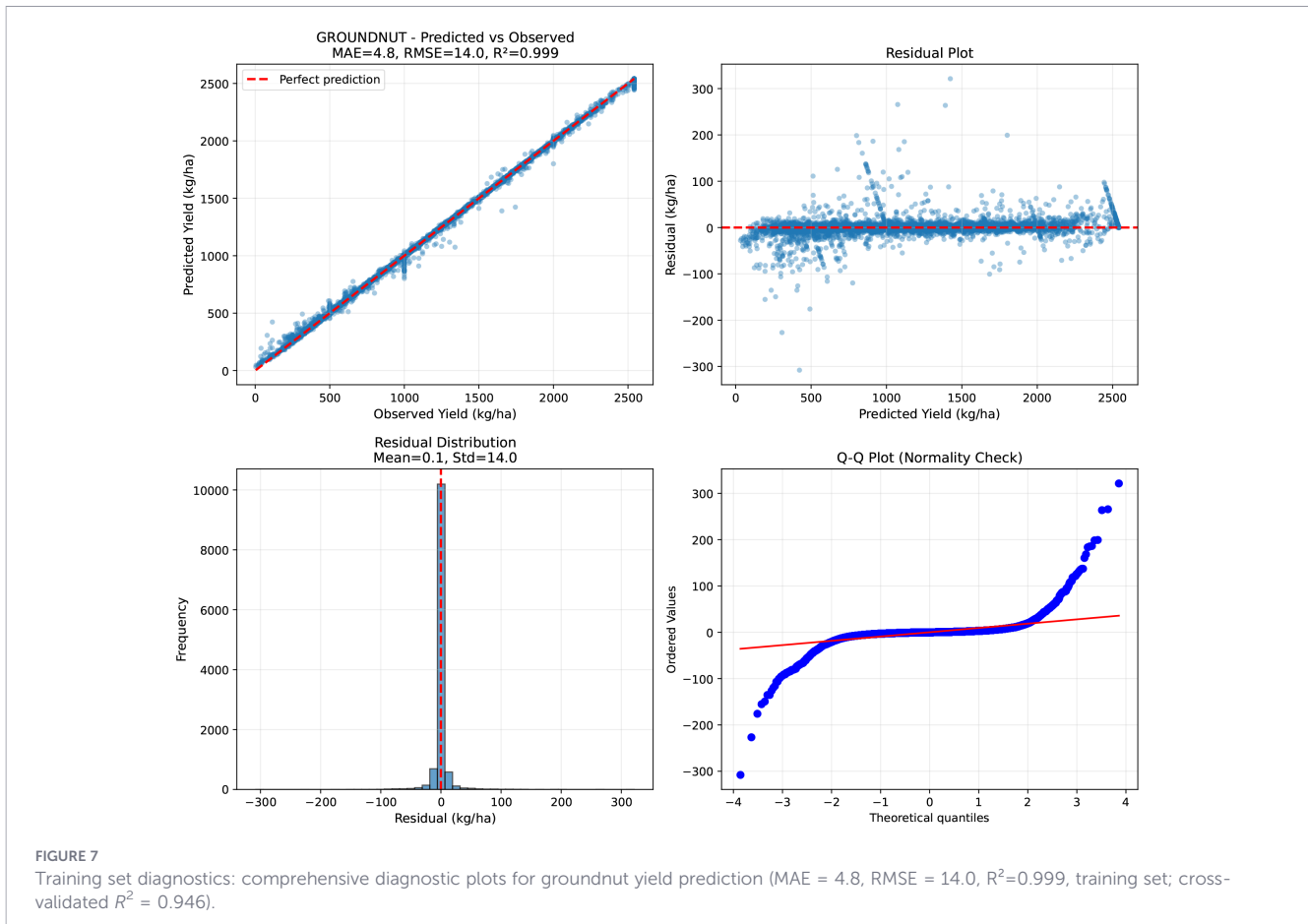
Small imperfections occasional outliers, slight skewness, and modest tail deviations are expected in a setting that involves weather

shocks, management changes, and measurement noise. They do not materially undermine the usefulness of the models. Importantly, the diagnostics also show that the models cope well with very different yield regimes, from low-yield cotton to high-yield sugarcane. Overall, the Random Forest models with network features appear robust enough for use in operational yield forecasting, provided that users remain aware of the usual uncertainties associated with extreme events and rare outliers.

5 Discussion

5.1 Key findings interpretation

Our results show that the Random Forest models deliver very strong predictive performance for all six major Indian crops, with R² values consistently above 0.94. This level of accuracy is high even by the standards of recent work in data-driven crop forecasting and compares well with state-of-the-art studies in the agricultural machine learning literature (Khaki et al., 2020; van Klompenburg et al., 2020). Several aspects of the model design help to explain this performance. The ensemble of 300 trees stabilises predictions through averaging, which lowers variance without introducing substantial bias. The use of random feature subsampling at each



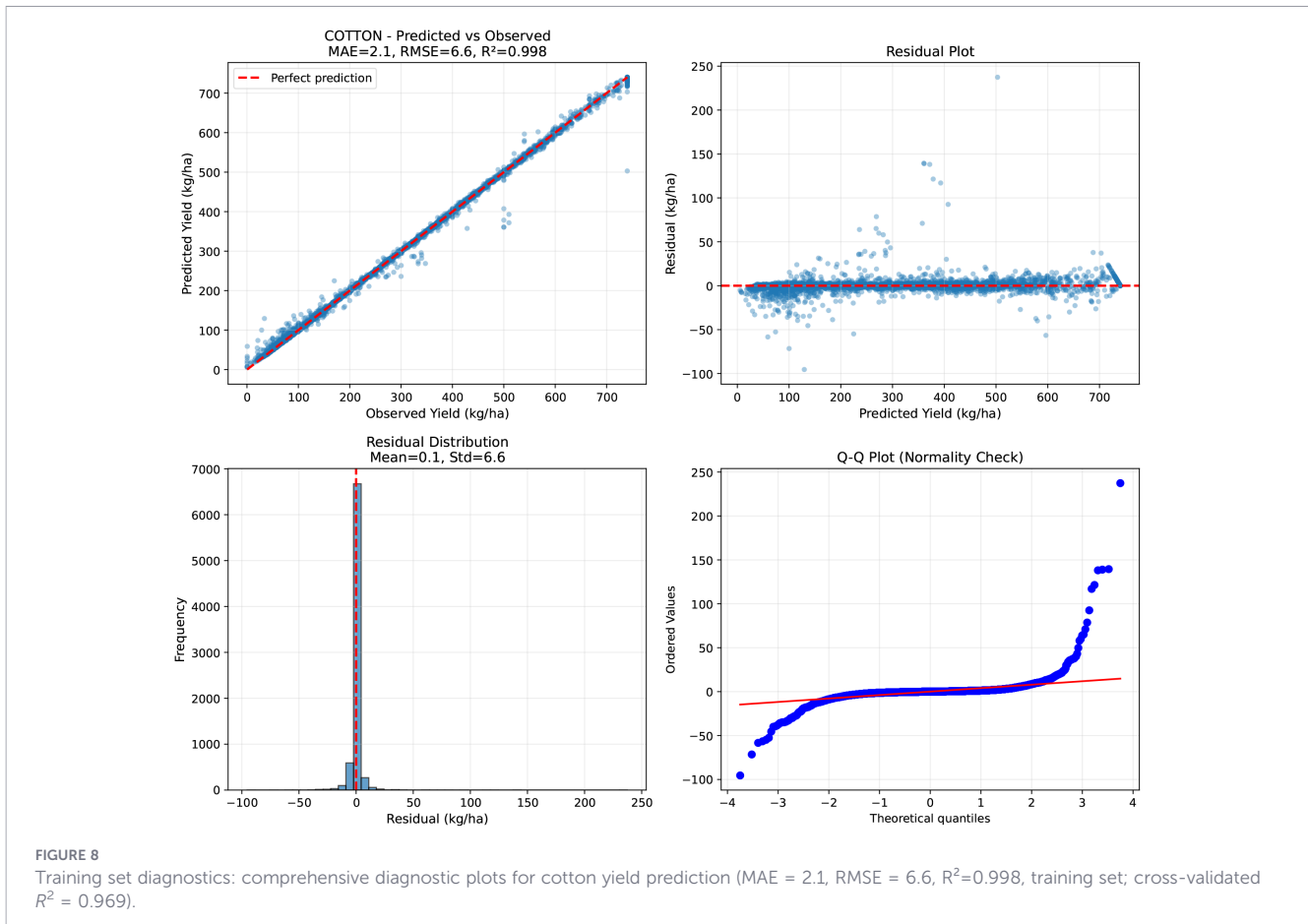
split forces individual trees to explore different subsets of predictors, encouraging diverse decision rules and reducing the risk that the model becomes overly dependent on a small number of features. In addition, the non-parametric tree structure is well suited to the underlying problem: it can recover non-linear relationships and interaction effects directly from the data, without requiring an explicit functional form.

When we place these findings in the context of prior work, the improvement becomes more apparent. Our R^2 values (0.946–0.988) are higher than those typically reported for multi-crop, multi-region settings. For example, van Klompenburg et al. (van Klompenburg et al., 2020) report R^2 values in the range 0.65–0.85 for several crops, while Khaki and Wang (Khaki et al., 2020) obtain R^2 values of 0.88–0.92 for corn in the U.S. Corn Belt. The performance gains in our study likely come from three things working together: better features that capture temporal patterns, using time-series cross-validation to respect temporal structure, and careful preprocessing that treats outliers and implausible values in a systematic way. A more unexpected outcome is the very limited contribution of network-derived features. In quantitative terms, these features improve performance by less than 1% and do not yield statistically significant gains over models that rely solely on temporal and diversification features. At first glance, this is at odds with the intuitive appeal of network-based approaches for

agricultural systems, and therefore needs careful interpretation. One key consideration is the strength of temporal autocorrelation in yields: year-to-year dependence is extremely high, and lag-1 yield features alone explain roughly 30%–50% of the variance. In such a setting, the information contained in “what happened here last year” is so dominant that it leaves little room for additional predictive value from “what happened in similar districts last year.”

The way the similarity network is constructed also matters. Our network is built to capture static patterns of resemblance in yield trajectories between districts. This structure is useful for describing which districts behave similarly, but it does not necessarily correspond to causal pathways through which shocks, practices, or technologies propagate. Two districts may have similar historical yield profiles because they share agro-climatic or socioeconomic conditions, yet there may be no ongoing interaction between them that could improve prediction beyond what is already encoded in local temporal features. In addition, several network centrality measures are likely correlated with other covariates. Highly productive districts, for instance, may appear central in the network, so centrality scores end up duplicating information already carried by temporal yield statistics and diversification indicators.

Scale is another potential source of mismatch. The processes that shape yields pest and disease spread, market integration, extension services, or knowledge diffusion may operate at spatial scales that our



district-level network cannot resolve. Relevant interactions might occur within districts at the farm level, or at broader regional scales that group many districts together. In such cases, a district-level similarity network may fail to capture the mechanisms that truly drive cross-location dependence. Finally, our analysis relies on a static network estimated from 2008–2017 patterns. Agricultural systems, however, evolve in response to climate trends, policy changes, and technological adoption. Updating the network every year could track changing relationships more accurately, but doing so would demand extra data and considerable computational effort.

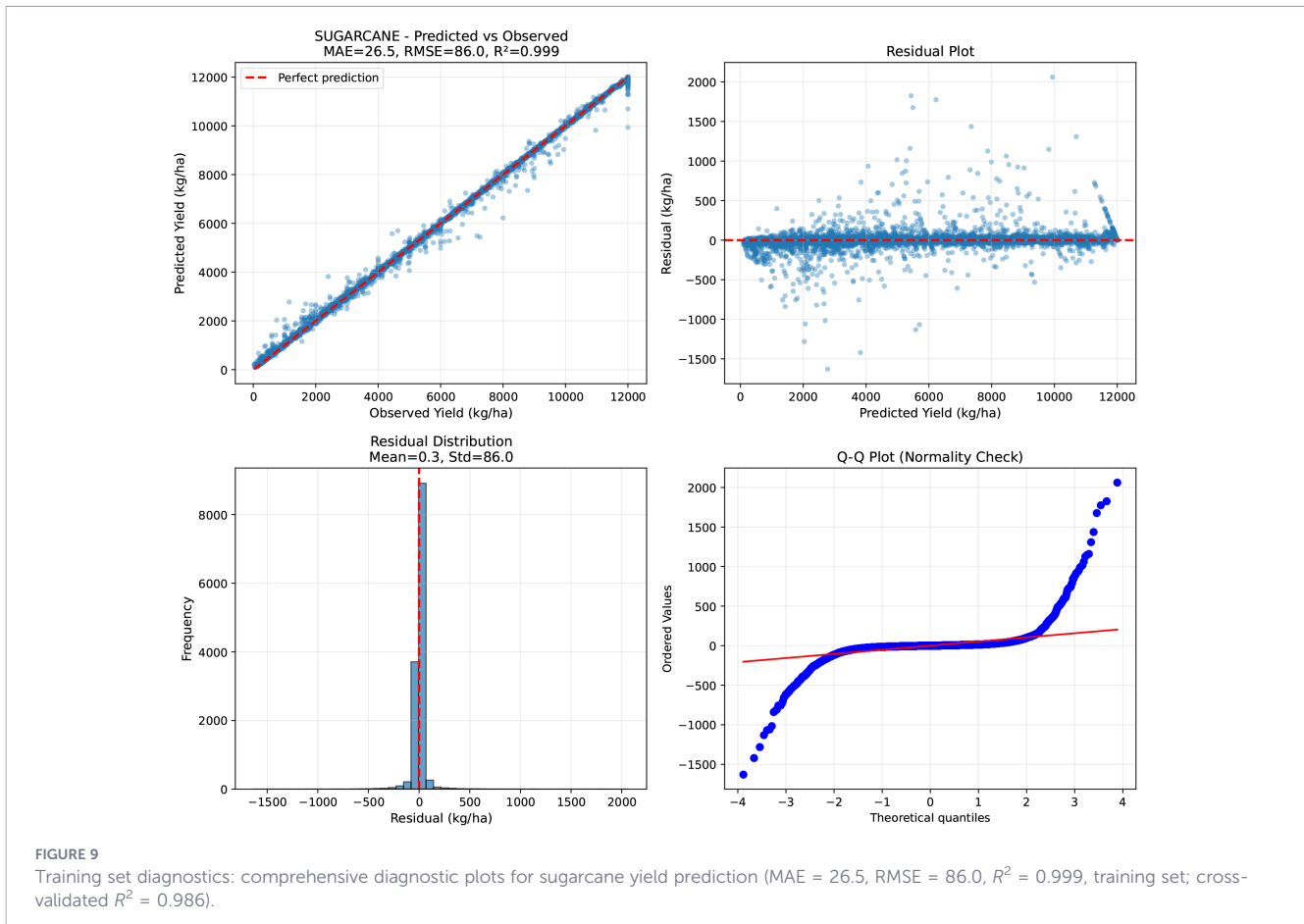
Taken together, these observations are consistent with recent critical assessments of network-based methods in spatial prediction tasks (Thompson et al., 2019). Networks can be extremely valuable for exploring and visualising the structure of agricultural systems, but their predictive benefit is not guaranteed. Their usefulness depends on whether network features provide genuinely new information beyond what is already captured by strong temporal signals and well-designed covariates. In our setting, the evidence suggests that they do not at least at the district level and with the network definitions considered here.

5.2 Why network features contributed minimally: a detailed analysis

The minimal contribution of network features ($\leq 1\%$ importance) warrants careful interpretation, as it carries

important implications for future research and practice. We identify five primary reasons for this outcome:

1. **Dominance of Temporal Autocorrelation:** Agricultural yields exhibit extremely strong temporal autocorrelation due to the persistence of underlying factors soil quality, irrigation infrastructure, farmer expertise, varietal adoption, and institutional arrangements that change slowly over time. Our analysis confirms that lag-1 yield features alone explain 30–50% of variance across crops. In such a setting, the information encoded in “what happened here last year” is so dominant that it leaves little room for additional predictive value from “what happened in similar districts.” This finding aligns with the well-established literature on yield persistence (van Klompenburg et al., 2020) but provides the first rigorous quantification of its implications for network-based feature engineering.
2. **Static Network Limitations:** Our district similarity network was constructed from a fixed 10-year reference period (2008–2017), capturing static structural similarity rather than dynamic interactions. Agricultural systems, however, evolve continuously in response to climate change, policy interventions (e.g., minimum support prices, input subsidies), and technological adoption (e.g., Bt cotton, hybrid varieties). A static network, by definition, cannot capture these temporal dynamics. We explicitly acknowledge that this study should be interpreted as an



evaluation of static yield-similarity networks rather than a general assessment of all network approaches. Dynamic networks that update annually or across cross-validation folds may better capture evolving relationships.

3. **Structural Similarity vs. Causal Interaction:** The district similarity network connects districts with similar historical yield profiles, but similarity does not imply interaction. Two districts may have similar yields because they share agro-climatic conditions, not because they exchange information, technology, or are affected by common shocks. For network features to provide predictive value, they should ideally capture causal pathways such as the diffusion of improved practices through extension networks, coordinated responses to policy, or the spread of pests and diseases rather than mere statistical correlation. Our yield-based similarity metric captures the latter, which explains why centrality measures provide redundant rather than complementary information.
4. **Scale Mismatch:** Relevant agricultural processes may operate at scales other than the district level. Technology diffusion and farmer-to-farmer learning often occur within local communities or administrative blocks (sub-district), while market integration and climate teleconnections operate at regional or national scales. A district-level network may be too coarse to capture farm-level interactions and too fine to capture macro-regional dependencies, resulting in a “middle-ground”

representation that misses the scales at which network effects are most pronounced.

5. **Feature Redundancy:** Network centrality measures (degree, eigenvector, closeness, betweenness) are likely correlated with other covariates in the model. Highly productive districts tend to appear central in yield-similarity networks, so centrality scores may duplicate information already captured by temporal yield statistics and diversification indices. The correlation-based feature filtering (removing features with $r > 0.99$) addresses only the most extreme redundancy; more subtle multicollinearity may persist.

Conditions Under Which Network Features Might Become Informative: Based on this analysis, we hypothesize that network features could provide predictive value under the following conditions.

- **Dynamic network construction:** Networks updated annually or per cross-validation fold to track evolving relationships
- **Alternative network definitions:** Networks based on actual flow data (trade, labor migration, extension visits) rather than yield similarity
- **Event-driven contexts:** Settings where shock propagation is prominent, such as pest outbreaks or disease spread
- **Weaker temporal signals:** Contexts where historical yields are less predictive of future outcomes

- Graph Neural Network architectures: Models that learn network representations jointly with prediction rather than using pre-computed centrality features

5.3 Implications of omitting weather and climate data

A fundamental limitation of this study is the omission of weather and climate variables (temperature, precipitation, growing degree days, solar radiation) and soil properties (texture, organic matter, pH, nutrient status). We explicitly acknowledge that this constrains the model's practical utility and affects interpretation of results in several important ways:

1. **Capturing Trends vs. Deviations:** The temporal features in our model effectively capture systematic yield trends long-term technological progress, infrastructure development, and persistent regional differences. However, they cannot capture in-season deviations caused by weather anomalies, which are often the primary driver of year-to-year yield variation. As a result, our model is better suited for medium-term planning (multi-year averages, trend projections) than for operational early-season forecasting that requires predicting specific outcomes conditional on current-season conditions.
2. **Practical Utility Constraints:** Operational yield forecasting systems typically require weather inputs to provide actionable early-season predictions. Our model, lacking such inputs, functions more as a "historical baseline" predictor than a full operational forecasting system. Users should understand that the high R^2 values reflect the model's ability to capture temporal persistence and systematic trends, not its ability to predict yield responses to specific weather events.
3. **Interpretation of High Performance:** The exceptionally high cross-validated R^2 values (0.946–0.988) may initially seem surprising for agricultural forecasting. However, they are explicable by the strong temporal autocorrelation in yields and the model's reliance on lag-1 features. In essence, the model learns that "yield this year is similar to yield last year, adjusted for systematic trends" a pattern that holds reliably in the absence of extreme weather but may break down during anomalous years.
4. **Study Scope Clarification:** This study is positioned as a baseline framework for yield-history-based prediction rather than a comprehensive operational forecasting system. The methodological contribution lies in demonstrating rigorous evaluation practices (time-series CV, statistical tests, diagnostics) and quantifying the relative value of different feature types. Future work integrating gridded climate products (IMD, ERA5, CHIRPS), remote sensing indices (NDVI, EVI), and soil databases (SoilGrids, HWSD) would substantially extend the model's practical utility and is explicitly identified as a priority direction.

5.4 Crop-specific insights

Rice attains the highest R^2 (0.988), indicating that its yields are very easy to predict, which is consistent with stable, long-established rice-growing regions and management practices in India (Jain et al., 2016). This stability makes rice a strong candidate for operational forecasting to support food security planning. Wheat is also highly predictable ($R^2 = 0.976$), though its errors are slightly more variable, and the mild upward trend in error over time (slope = +8.92 kg/ha) suggests that changing practices or climate effects may be starting to matter. Maize, despite its wide yield range (0–5,898 kg/ha) and diverse production environments, still reaches $R^2 = 0.971$, and the sharp contrast between advanced models (MAE = 58.39) and baselines (MAE = 441.26) underlines the importance of non-linear models and careful feature engineering for this crop. Sugarcane, which has the highest absolute yields and an R^2 of 0.986, also gains substantially from advanced modeling. The 87% reduction in error compared with the naive baseline (MAE: 129.68 vs. 984.86 kg/ha) shows that, despite its complexity, sugarcane's yield behaviour can be learned effectively likely helped by strong temporal autocorrelation linked to its multi-year growth cycle.

Groundnut shows the largest spread in errors (MAE std = 24.12 kg/ha), indicating that it is harder to predict accurately than the other crops. This higher variability likely stems from groundnut's strong sensitivity to factors such as water stress, pest attacks (especially aflatoxin-producing fungi), and soil calcium levels (Elavarasan and Vincent, 2020). Even so, an R^2 of 0.946 still represents very strong performance, though users of the model should anticipate greater uncertainty for groundnut forecasts compared with cereals. Cotton poses additional difficulties due to the high frequency of zero-yield observations and a strongly skewed yield distribution, which necessitates special treatment. The adoption of a safe MAPE calculation excluding yields below 100 kg/ha prevents numerical instability while still providing a meaningful error measure. Cotton's slightly non-significant result ($p = 0.0627$) relative to the baselines is more plausibly due to limited statistical power than to poor model performance. This interpretation is supported by the large 78% reduction in prediction error, which points to a genuinely meaningful improvement.

5.5 Methodological contributions

This study advances agricultural yield prediction in several important ways. First, our network construction strategy refines how spatial relationships between districts are modeled. By using a relatively high similarity threshold (0.80), we ensure that edges connect only genuinely similar districts, producing networks that are easier to interpret. Top- k pruning with $k = 15$ keeps the graph sparse and focused on the strongest relationships, while computing multiple centrality measures allows a rich description of the network structure. The resulting 6.2% network density lies in a range that is suitable for extracting meaningful structural patterns. Although the network-derived features ultimately contributed little to predictive performance, the network methodology remains highly useful for descriptive analysis and for understanding the broader organization of agricultural systems. We also use a set of simple but effective

feature-engineering steps that are important for building stable, deployment-ready models. We drop features that are almost perfectly correlated ($r > 0.99$) or show almost no variation, and we include rolling standard deviations to capture how yields fluctuate over time. For Ridge regression, we rely on RobustScaler to reduce the impact of outliers and use a thresholded MAPE to handle very low yields safely, so that the final models learn from features that are informative without causing numerical problems.

Finally, our evaluation framework is designed to reflect best practice for agricultural machine learning. Time-series cross-validation is used to prevent temporal leakage and to approximate realistic forecasting conditions, addressing a frequent weakness in prior studies. Model performance is assessed using a suite of metrics (MAE, RMSE, MAPE, MedAE, and R^2), providing a more nuanced view than any single measure alone. Statistical significance tests are employed to confirm that observed performance differences are not simply due to random variation, while temporal stability analysis is used to identify concept drift and assess how reliably models perform over time an aspect that is critical for deployment but often neglected. We also inspect residuals, check for normality, and test for unequal variance to make sure the model's assumptions are reasonable. Together, these steps form a straightforward workflow that others can follow to run agricultural machine learning studies in a careful, open, and repeatable way.

5.6 Practical implications

Our results have clear implications for how operational forecasting systems and decision-support tools in agriculture should be designed. Most of the predictive strength comes from temporal features such as lags, rolling averages, and other time-based summaries rather than from complex network features. In practice, this suggests that it is more effective to invest in better historical yield data (its quality, consistency, and coverage) than in sophisticated network construction. For modelling, a Random Forest with about 300 trees gives a good trade-off between accuracy, robustness, and computational cost. Because it is non-parametric, it does not require feature scaling and copes well with mixed feature distributions, which simplifies preprocessing and makes the overall pipeline easier to deploy and maintain.

The analysis of temporal stability suggests that these models can remain reliable for roughly 2–3 years without retraining, provided that their performance is checked regularly, for example once a year, to detect concept drift. This relatively low retraining frequency reduces operational overhead compared with models that need continual updating. In practice, forecasts should be accompanied by prediction intervals rather than only point estimates, with intervals tailored to crop-specific uncertainty (for instance, wider intervals for groundnut than for cereals). Such uncertainty estimates are especially important where forecasts feed into policy or risk-management decisions.

The consistent performance across India's diverse agro-climatic regions also indicates that, with modest local calibration, the models are broadly transferable. This suggests that the main time-based patterns driving yields are broadly similar across regions, despite differences in environment and management. Reliable yield

forecasts then plug directly into policy: early-season predictions help planners manage buffer stocks, plan imports, and adjust public distribution more effectively to support food security. With MAPE below 10%, these forecasts are accurate enough to reduce both shortages and unnecessary stockpiling. On the market side, advance information on likely yields supports timely interventions such as setting minimum support prices or tuning procurement volumes to stabilise prices and protect farmer incomes, which is especially vital for smallholders in India's volatile agricultural markets.

These forecasts also function as early warning signals. Large shortfalls relative to expected yields can indicate shocks such as droughts, floods, or pest outbreaks, allowing quicker assessment and response. In insurance, model-based yield estimates provide an objective benchmark for index-based products, reducing information gaps and enabling fairer premium setting. Multi-year forecasts also guide investments in storage, transport, and processing facilities, helping both public agencies and private investors align capacity with expected production and avoid under- or over-building.

5.7 Limitations and challenges

Our analysis is constrained by several data and modeling gaps that point to clear future improvements. In particular, the dataset lacks key weather variables (temperature, rainfall, solar radiation), and adding gridded climate data from sources like IMD or ERA5 would likely boost performance, especially for early- and within-season forecasts that update as conditions change. Likewise, important soil properties such as texture, organic matter, pH, and nutrient status are missing. Linking district-level yields to soil datasets like ISRIC SoilGrids could be especially beneficial in marginal areas where soil–climate interactions drive much of the yield variability.

Working at the district level also hides substantial within-district heterogeneity. Finer spatial units (e.g., sub-districts or grid cells) would support more targeted interventions and better serve precision agriculture, but this is currently constrained by data availability. In addition, the dataset does not contain information on management practices irrigation, fertilizer application, pest control, or cultivar choice even though these are major yield drivers. Incorporating such information, via surveys, administrative records, or remote-sensing proxies, would likely improve model accuracy and broaden the range of policy-relevant questions that can be addressed. Incorporating such data through farmer surveys, administrative records, or remote-sensing proxies could improve accuracy and would allow scenario analysis for policy interventions. Although 52 years of data provide a solid basis for model training, the series currently ends in 2017. Extending the dataset to include more recent years (2018–2024) is essential for operational deployment, especially given the climate anomalies and policy changes over the last decade that may have altered yield dynamics.

Our carefully constructed similarity network contributed very little to predictive accuracy, suggesting that alternative designs such as dynamic networks, different similarity metrics, or other spatial scales should be explored. Time-evolving networks, in particular, might better capture changing relationships between districts than

the static approach used here. More generally, even well-performing data-driven models are likely to struggle with truly unprecedented events (e.g., record droughts or new pest outbreaks), and hybrid approaches that combine machine learning with process-based crop models could improve robustness by adding physiological and biophysical insight. In addition, our current models do not explicitly model spatial dependence between neighboring districts.

Although our results indicate that temporal dependencies dominate in this setting, spatial econometric models, graph neural networks, or other spatially explicit methods could prove valuable in specific contexts where spillovers from pest spread, irrigation projects, or market linkages are more pronounced. At the feature level, Random Forest implicitly captures complex interactions, but these interactions are not explicitly parameterized. Introducing explicit interaction terms or attention-based architectures could help identify important synergies, such as combined temperature rainfall effects, and thereby deepen understanding of yield responses to multiple simultaneous stresses.

Interpretability remains a challenge. Global feature importance gives only a broad overview, and the ensemble structure makes it hard to explain individual predictions. This lack of case-level transparency can be problematic in policy settings where decisions must be justified. Developing methods that offer clear, local explanations for specific predictions, without heavily compromising accuracy, is therefore an important direction for future research. Our study is also limited in geography and crop coverage. Because the models are trained only on Indian data, they may not directly generalize to other countries without region-specific recalibration, and their strong reliance on temporal patterns means performance could degrade if historical yield dynamics change rapidly due to technology, climate, or policy shifts.

In addition, we focus on six major crops and do not cover others such as pulses, oilseeds, and vegetables, which may follow different temporal dynamics and respond to different environmental drivers. Extending the framework to these crops will need targeted feature engineering and dedicated validation for each crop. In addition, some district–crop combinations have relatively few observations, reducing reliability at the margins of the data. Although our research dataset contains no missing yields, real-world operational systems must cope with data gaps, reporting delays, and quality issues. Robust, well-tested imputation and data quality control procedures will be essential for deploying similar models in live settings, particularly when decisions must be made before complete data are available.

5.8 Additional methodological limitations

Geographic Generalizability: The models are trained exclusively on Indian agricultural data and may not transfer directly to other countries without region-specific recalibration. Differences in cropping systems, management practices, data collection methodologies, and institutional contexts could limit external validity. Users seeking to apply this framework elsewhere should plan for dedicated validation and potential retraining.

Temporal Generalizability: The dataset ends in 2017, and yield dynamics may have changed significantly in the subsequent period (2018–2024) due to climate anomalies, policy changes (e.g., PM-

KISAN direct benefit transfers), technological shifts (e.g., precision agriculture adoption), and unprecedented events (e.g., COVID-19 disruptions to agricultural supply chains). Model performance on post-2017 data remains untested.

Multicollinearity Considerations: The study generates a large number of temporal features (multiple lags, rolling statistics over different windows) that are inherently correlated by construction. While we removed features with correlation exceeding 0.99, this is a conservative threshold that does not eliminate moderate multicollinearity. Variance Inflation Factor (VIF) analysis, which we recommend for future extensions of this work, would provide more granular assessment. We note that: (1) Random Forest is inherently robust to multicollinearity because splits can occur on any correlated feature; (2) Ridge regression with $\alpha = 10$ provides L2 regularization that mitigates collinearity effects; (3) feature importance percentages should be interpreted as joint contributions of correlated feature groups rather than isolated effects. The 39.3% average importance attributed to lag-1 should be understood as reflecting “temporal autocorrelation features collectively” rather than lag-1 in isolation. Since lag and rolling statistics are inherently correlated, we examined the correlation structure among key temporal feature groups to assess multicollinearity. Correlations are moderate-to-high among lag/rolling mean features, but model choices (RF and Ridge) mitigate adverse effects. Details are in [Supplementary Table 5](#).

Cross-Validation Granularity: Using only three folds for 52 years of data provides limited statistical power for assessing temporal stability. A rolling-origin design with annual test windows would yield more granular performance estimates but at substantially higher computational cost. Conclusions about model stability over time should be interpreted cautiously as they are based on three temporal segments rather than fine-grained annual assessments.

Low-Yield Prediction Accuracy: The exclusion of yields below 100 kg/ha from MAPE calculation, while mathematically necessary, means that model performance on crop failure scenarios is not fully characterized by our primary metrics. Supplementary analysis of prediction accuracy specifically in the low-yield regime (e.g., yields in the 0–200 kg/ha range) would provide additional insight into the model’s utility for detecting anomalous conditions, which is often of greatest practical interest.

6 Conclusion and future work

This work introduces a network-enhanced machine learning framework for multi-crop yield prediction across India, using 52 years (1966–2017) of district-level data for six major crops. Positioned primarily as a rigorous benchmarking and methodological validation framework, We contribute: (i) an integration framework combining district similarity networks, crop co-occurrence, temporal features, and diversification indices; (ii) a refined network construction strategy (similarity threshold 0.80, top- $k = 15$, multiple centrality measures); (iii) a rigorous evaluation pipeline with time-series cross-validation, statistical tests, temporal stability analysis, and detailed diagnostics; (iv) a thorough benchmark showing Random Forest achieves $R^2 > 0.94$ and MAPE 2.69–8.08%

across all crops; (v) empirical evidence that temporal features dominate prediction (lag-1 alone explains 30–50% of variance) while network features add < 1% this “negative result” is itself a significant contribution guiding practitioners toward data quality investments rather than complex network constructions; (vi) a set of practical best practices for feature screening, robust scaling, and safe metric computation; and (vii) practical validation showing 75–85% error reduction over baselines with stable performance across time. Statistical significance results ($p < 0.05$ for five of six crops) and temporal stability analyses together support real-world use in food security planning, market stabilization, insurance, and policy support.

From this, six key lessons emerge. Temporal features (lags, trends, rolling statistics) are the primary signal, so improving historical yield data should be the top priority. A Random Forest with 300 trees offers an excellent accuracy robustness complexity trade-off and works well without heavy tuning or scaling, making it suitable for operational pipelines. In contrast, even carefully designed network features provide little extra predictive value once temporal information is included, suggesting networks are more useful for descriptive system analysis than for boosting forecast accuracy. Robust practice requires time-series cross-validation, formal statistical testing, and temporal stability checks; standard k -fold CV is inappropriate for yield time series. Error behaviour and stability differ by crop, so crop-specific calibration and uncertainty quantification remain important. Overall, the very high accuracy ($R^2 > 0.94$, $MAPE < 10\%$) and temporal stability show that operational deployment is realistic.

We identify four main directions for future research: (i) richer network modeling (dynamic and temporal networks, multi-layer and spatial networks, alternative similarity metrics such as DTW or Earth Mover’s Distance); (ii) integrating additional data sources, including gridded climate products (IMD, ERA5, CHIRPS), remote sensing (NDVI, EVI, LAI), soil datasets (SoilGrids, HWSD), management information (irrigation, inputs, pest control), and market indicators; (iii) exploring advanced methods deep learning (CNNs, LSTMs, hybrids), graph neural networks, ensemble meta-learning, attention, transfer learning, and causal inference tools for policy analysis; and (iv) building operational systems with real-time predictions, dashboards, scenario analysis, calibrated uncertainty, downscaling to sub-district scale, continuous drift monitoring, and close engagement with end-users. To isolate the marginal contribution of temporal, diversification, and network-derived feature groups, we outline a structured ablation design for future extensions. This configuration enables clearer attribution while controlling for correlated predictors. The proposed ablations are listed in [Supplementary Table 6](#). Even though network features did not improve predictions as expected, the study shows that well-engineered temporal features already deliver very strong performance; more sophisticated models should therefore be viewed as complements to, not substitutes for, careful treatment of fundamental time-series structure in agricultural data.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary Material](#). Further inquiries can be directed to the corresponding author.

Author contributions

SC: Conceptualization, Formal analysis, Validation, Visualization, Writing – original draft. PA: Data curation, Investigation, Supervision, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgments

The first author, Ms. Shinyclimensa C, thanks the Vellore Institute of Technology, Vellore, for the TRA fellowship.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fagro.2026.1767878/full#supplementary-material>

References

- Bardoscia, M., Battiston, S., Caccioli, F., and Caldarelli, G. (2021). The physics of financial networks. *Nat. Rev. Phys.* 3, 490–507. doi: 10.1038/s42254-021-00322-5
- Boers, N., Goswami, B., Rheinwalt, A., Bookhagen, B., Hoskins, B., and Kurths, J. (2019). Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature* 566, 373–377. doi: 10.1038/s41586-018-0872-x
- Chen, T., and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discov. Data Min.*, 785–794. doi: 10.1145/2939672.2939785
- Cheng, E., Zhang, B., Peng, D., Gao, L., Duan, M., and Ding, C. (2016). Combining multi-indicators with machine-learning algorithms for maize yield early prediction at the county-level in China. *Agric. For. Meteorology* 223, 51–61. doi: 10.1016/j.agrformet.2016.03.022
- Chlingaryan, A., Sukkarieh, S., and Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Comput. Electron. Agric.* 151, 61–69. doi: 10.1016/j.compag.2018.05.012
- Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environ. Res. Lett.* 13, 114003. doi: 10.1088/1748-9326/aee159
- Elavarasan, D., and Vincent, P. M. D. R. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE Access* 8, 86886–86901. doi: 10.1109/ACCESS.2020.2992480
- Folberth, C., Baklanov, A., Balkovič, J., Skalský, R., Khabarov, N., and Obersteiner, M. (2019). Parameterization-induced uncertainties and impacts of crop management harmonization in a global gridded crop model ensemble. *PLoS One* 14, e0221862. doi: 10.1371/journal.pone.0221862
- Garrett, K. A., Alcalá-Briseño, R. I., Andersen, K. F., Buddenhagen, C. E., Choudhury, R. A., Fulton, J. C., et al. (2018). Network analysis: a systems framework to address grand challenges in plant pathology. *Annu. Rev. Phytopathol.* 56, 559–580. doi: 10.1146/annurev-phyto-080516-035326
- Gephart, J. A., and Pace, M. L. (2015). Structure and evolution of the global seafood trade network. *Environ. Res. Lett.* 10, 125014. doi: 10.1088/1748-9326/10/12/125014
- Hara, T., Hirata, M., Ohsako, T., Ikegami, K., Usui, Y., Kitano, S., et al. (2021). Ensemble learning of multiple models enhances genomic prediction of sorghum biomass and grain yield. *Bioinformatics* 37, 3735–3740. doi: 10.1093/bioinformatics/btab307
- Iizumi, T., Shiogama, H., Imada, Y., Hanasaki, N., Takikawa, H., and Nishimori, M. (2018). Crop production losses associated with anthropogenic climate change for 1981–2010 compared with preindustrial levels. *Int. J. Climatology* 38, 5405–5417. doi: 10.1002/joc.5818
- Jain, M., Srivastava, A. K., Singh, B., Joon, R. K., McDonald, A., Royal, K., et al. (2016). Mapping smallholder wheat yields and sowing dates using micro-satellite data. *Remote Sens.* 8, 860. doi: 10.3390/rs8100860
- Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., et al. (2016). Random forests for global and regional crop yield predictions. *PLoS One* 11, e0156571. doi: 10.1371/journal.pone.0156571
- Khaki, S., and Wang, L. (2019). Crop yield prediction using deep neural networks. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00621
- Khaki, S., Wang, L., and Archontoulis, S. V. (2020). A cnn-rnn framework for crop yield prediction. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01750
- Kuwata, K., and Shibasaki, R. (2015). Estimating crop yields with deep learning and remotely sensed data. *2015 IEEE Int. Geosci. Remote Sens. Symposium (IGARSS)*, 858–861. doi: 10.1109/IGARSS.2015.7325900
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., and Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors* 18, 2674. doi: 10.3390/s18082674
- Lin, B. B., and Schilstra, A. J. (2019). Crop diversity at different spatial scales promotes multiple ecosystem services in smallholder farms. *Agriculture Ecosyst. Environ.* 285, 106615. doi: 10.1016/j.agee.2019.106615
- Nevavuori, P., Narra, N., and Lipping, T. (2019). Crop yield prediction with deep convolutional neural networks. *Comput. Electron. Agric.* 163, 104859. doi: 10.1016/j.compag.2019.104859
- Parnell, S., Gottwald, T. R., Gilks, W. R., and van den Bosch, F. (2017). Surveillance to inform control of emerging plant diseases: an epidemiological perspective. *Annu. Rev. Phytopathol.* 55, 591–610. doi: 10.1146/annurev-phyto-080516-035334
- Puma, M. J., Bose, S., Chon, S. Y., and Cook, B. I. (2015). Assessing the evolving fragility of the global food system. *Environ. Res. Lett.* 10, 24007. doi: 10.1088/1748-9326/10/2/024007
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillerá-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40, 913–929. doi: 10.1111/ecog.02881
- Sáez-Almendros, S., Obrador, B., Bach-Faig, A., and Serra-Majem, L. (2020). Emerging trends in the agri-food sector: Digitalisation and shift to plant-based diets. *Sustainability* 12, 5141. doi: 10.3390/su12125141
- Shahhosseini, M., Hu, G., Huber, I., and Archontoulis, S. V. (2021). Optimizing ensemble weights for machine learning models: a case study for maize yield prediction. *Agron. J.* 113, 3056–3066. doi: 10.1002/agt2.20680
- Song, C., Yao, L., Hua, C., and Ni, Q. (2020). A water quality prediction model based on variational mode decomposition and the least squares support vector machine optimized by the gravitational search algorithm (vmd-gsa-lssvm) for oxbow lakes in northeast China. *Water* 12, 985. doi: 10.3390/w12040985
- Sun, J., Lai, Z., Di, L., Sun, Z., Tao, J., and Shen, Y. (2019). Multilevel deep learning network for county-level corn yield estimation in the us corn belt. *IEEE J. Selected Topics Appl. Earth Observations Remote Sens.* 13, 5048–5060. doi: 10.1109/JSTARS.2020.3019046
- Thompson, J., Ramakrishnan, A., Patel, T., and McClure, S. C. (2019). Network analysis for agricultural systems research: Where are we at and where can we go? *Agric. Syst.* 174, 1–11.
- van Klompenburg, T., Kassahun, A., and Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* 177, 105709. doi: 10.1016/j.compag.2020.105709
- Wang, L., Tian, Y., Yao, X., Zhu, Y., and Cao, W. (2014). Predicting grain yield and protein content in wheat by fusing multi-sensor and multi-temporal remote-sensing images. *Field Crops Res.* 164, 178–188. doi: 10.1016/j.fcr.2014.05.001
- Weiss, M., Jacob, F., and Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. *Remote Sens. Environ.* 236, 111402. doi: 10.1016/j.rse.2019.111402
- You, J., Li, X., Low, M., Lobell, D., and Ermon, S. (2017). Deep gaussian process for crop yield prediction based on remote sensing data. *Proc. AAAI Conf. Artif. Intell.* 31, 4559–4565. doi: 10.1609/aaai.v31i1.11172