AI

MDPI

Article

A Counterfactual-Enabled Agricultural Decision Support Framework for Sustainability-Aware Groundnut Yield Prediction Using Bayesian-Optimized XGBoost

Rekha R Nair [1,2], Tina Babu [1], Sumendra Yogarayan [2,3,*] and Siti Fatimah Binti Abdul Razak [2,3]

[1] Department of Computer Science and Engineering, Alliance School of Advanced Computing, Alliance University, Bengaluru 562106, India; rekha.nair@alliance.edu.in (R.R.N.); tina.babu@alliance.edu.in (T.B.)
[2] Faculty of Information Science and Technology (FIST), Multimedia University, Jalan Ayer Keroh Lama, Bukit Beruang 75450, Malaysia
[3] Centre for Intelligent Cloud Computing, COE for Advanced Cloud, Multimedia University, Melaka 75450, Malaysia; fatimah.razak@mmu.edu.my
* Correspondence: sumendra@mmu.edu.my

Abstract

Sustainable agricultural planning requires predictive frameworks that can capture spatiotemporal variability, sustainability dynamics, and the potential outcomes of alternative management scenarios. The research proposes TSAFI-DT, a retrospectively validated, data-driven Digital Twin prototype integrating spatiotemporal data reconstruction, sustainability-state representation, hierarchical yield forecasting, counterfactual analysis, and scenario simulation. The framework operates on historical district-level APY observations and therefore represents a retrospective approximation of Digital Twin operation rather than a continuously synchronized cyber-physical agricultural Digital Twin. The Extended Regenerative Agriculture Index (eRAI) combines crop diversity, productivity–stability, land-use efficiency, and yield-trend information to characterize district-level sustainability states. A Bayesian-optimized XGBoost model is employed for one-step-ahead yield forecasting under temporal validation, while fixed-effects and synthetic-control analyses provide complementary associational and intervention-associated evidence. Evaluation using district-level groundnut data from India during 1997–2023 demonstrates that the proposed predictor achieves an RMSE of 0.171 t/ha and $R^2 = 0.92$, outperforming the evaluated baselines with statistically significant differences ($p < 0.05$). The fully adjusted fixed-effects model identifies a positive association between higher sustainability states and yield, while retrospective Digital Twin replay demonstrates close temporal agreement between predicted and observed outcomes. Model-based scenario simulations indicate predicted yield increases of up to 12.4% under the evaluated sustainability-state perturbations; these estimates represent counterfactual sensitivity rather than guaranteed causal effects. TSAFI-DT provides a reproducible framework for sustainability-aware agricultural forecasting, comparative scenario exploration, and data-driven decision support.

Keywords: digital twin system; explainable artificial intelligence (XAI); spatiotemporal data analytics; agricultural crop-yield prediction; counterfactual analysis

Check for updates

Academic Editor: Miguel Angel Cazorla

Received: 24 July 2026
Revised: 2 September 2026
Accepted: 7 September 2026
Published: 14 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The worldwide agricultural industry is in the midst of a model shift spurred by the rapid integration of artificial intelligence, information sciences, big data analytics,

and cyber-physical systems. Innovations in these fields have created new potential for agricultural resource utilization, mostly in the face of growing demands for sustainable food production, global climate variability, aggressive soil depletion, and limited natural resources [1,2]. While conventional agricultural modelling has heavily depended upon fixed statistical modelling techniques and trend analysis over time, they do not effectively account for the interaction within the spatiotemporal domain of the relevant agricultural system [3,4].

The Digital Twin (DT) technology has emerged as a revolutionary approach capable of resolving these issues, as it allows the creation of a virtual replica of physical systems [5–7]. A Digital Twin constructs a virtual representation of a physical system by combining live and historical information with computational modelling to replicate its operational dynamics, with the ability to perform predictive analysis, explore different scenarios, and aid decision-making [8,9]. In agriculture, DT enables fusion among diverse data sources like Internet of Things (IoT) sensor networks, remote sensing data, climate data, and crop management data to allow a holistic understanding of agricultural systems [10–14].

Recently, several works have demonstrated the potential benefits of applying the Digital Twin context in agriculture to improve yields and sustainability [15–17]. Recent related work has also explored regenerative-agriculture-oriented field intelligence for sustainable groundnut yield prediction using sustainability indicators, XGBoost, and explainable AI [18]. For example, DT-based crop type identification systems were proposed by combining deep learning solutions with remote sensing data and were measured to be a more accurate and scalable setup [19,20]. Also, IoT-enabled DT system architectures have been designed for precision irrigation to promote resource-efficient solutions, as well as real-time monitoring of plant growth status to tackle the shortage of water resources and waste of fertilizers [11,21]. However, AI-powered analytical modules in the DT context have shown themselves to be useful for related activities, including pest/disease detection, earth/soil condition analysis, climate-controlled and adaptive planting strategies [22–26].

But there are still several challenges that prevent the general application of Digital Twins in agriculture. Agricultural data typically face gaps, heterogeneity, or other quality issues, thereby making the data sparse and inconsistent [2,3]. In addition, most of the applications of Digital Twins are limited to prediction models without causality in the models or the integration of interventions, which impedes the evaluation of policies or environmental changes [27,28]. An additional significant limitation is the poor inclusion of sustainability indicators in the design of Digital Twin structures. For the field of sustainable agriculture, a comprehensive but interconnected analysis should consider indicators like soil health, land-use efficiency, and plant biodiversity, but these are not all captured by the DT models [11,21]. Also, the lack of closed-loop feedback integration inhibits adaptation and learning processes in many Digital Twin solutions [29,30]. Recent studies also suggest that simulation and scenario-based analysis should be built into the DT system for 'what-if' analysis [31,32]. This allows decision-makers to know in advance the impact of different intervention measures, like new agricultural practices or policy tools. For the causal analysis, the integration of causal inference techniques into DT systems was recognized as another requirement to better explore the cause–effect relationships among factors' inputs and agricultural output [22,33].

To address these limitations, TSAFI-DT integrates spatiotemporal reconstruction, sustainability-state characterization, hierarchical predictive learning, counterfactual and fixed-effects analyses, retrospective state updating, and model-based scenario exploration within a unified data-driven framework.

The primary objective of this study is to develop TSAFI-DT, a data-driven Digital Twin framework for sustainability-aware agricultural yield forecasting, temporal state charac-

terization, and scenario-based decision support using historical district-level agricultural data. TSAFI-DT is positioned in the research as a retrospective, data-driven Digital Twin prototype whose virtual state is reconstructed and sequentially updated from historical observations, rather than as a real-time cyber-physical Digital Twin. The specific objectives are as follows:

- To construct a temporally consistent district-level agricultural data representation through spatiotemporal synchronization, tiered missing-data reconstruction, and structural-change detection while preventing future-information leakage.
- To develop the Extended Regenerative Agriculture Index (eRAI) by integrating crop diversity, productivity stability, land-use response, and yield-trend dynamics for comparative characterization of district-level sustainability states.
- To model the temporal evolution and persistence of sustainability conditions using state-transition analysis and identify heterogeneous sustainability trajectories across agricultural districts.
- To develop a hierarchical Bayesian-optimized XGBoost forecasting model that integrates historical, spatial, seasonal, and sustainability-state information for accurate one-step-ahead crop-yield prediction under time-aware validation.
- To incorporate counterfactual and fixed-effects panel analyses for examining intervention-associated evidence and adjusted relationships between sustainability states and agricultural productivity without assuming unconditional causal identification.
- To enable model-based "what-if" analysis by simulating predefined changes in virtual sustainability-state variables and quantifying their associated sensitivity in predicted agricultural yield.
- To implement retrospective sequential state updating as an approximation of the Digital Twin feedback cycle and evaluate the temporal fidelity of the virtual agricultural representation using historically observed outcomes.
- To evaluate the proposed framework through predictive benchmarking, statistical significance testing, model interpretability, sustainability-state analysis, scenario simulation, and predictive ablation to establish its utility for sustainability-aware agricultural decision support.

The remaining part of this paper is organized as follows. Section 2 presents the proposed TSAFI-DT methodology, including spatiotemporal data reconstruction, structural-change detection, sustainability-state representation, hierarchical yield forecasting, counterfactual and fixed-effects panel analysis, Digital Twin scenario simulation, and the experimental evaluation protocol. Section 3 presents the Results and Discussion, covering reconstruction performance, sustainability dynamics, predictive evaluation, counterfactual and fixed-effects analyses, scenario simulation, retrospective temporal validation, and ablation analysis. Finally, Section 4 concludes the study by summarizing the main findings, limitations, and directions for future research.

2. Materials and Methods

TSAFI-DT is a retrospectively validated, data-driven Digital Twin prototype designed for district-level agricultural yield forecasting, sustainability-state characterization, counterfactual analysis, and model-based scenario exploration. The framework integrates spatiotemporal data reconstruction, structural-change detection, sustainability-state representation, hierarchical yield prediction, explainable artificial intelligence, counterfactual and fixed-effects analyses, scenario simulation, and retrospective sequential state updating.

The overall workflow of TSAFI-DT is illustrated in Figure 1. The framework begins with the reconstruction of a temporally consistent district-year agricultural dataset from

heterogeneous climate, soil, agricultural, spatial, and temporal observations. The reconstructed data are then used to construct the sustainability-state representation through CDI, SHP, LUE, and YTM, which are integrated into the eRAI composite score. Structural-change detection and sustainability-state transition analysis are subsequently used to characterize temporal regime shifts, state persistence, and transitions across districts. The resulting sustainability information is incorporated into the HB-XGBoost model for one-step-ahead yield prediction, while SHAP analysis is used to quantify the contribution of the model inputs to the predicted yield. Counterfactual and fixed-effects analyses provide complementary evidence on the associations between sustainability-related changes and yield outcomes. The Digital Twin scenario layer then performs controlled perturbations of selected sustainability-state variables and evaluates the resulting district-level and aggregate model responses. Finally, retrospective sequential replay updates the virtual state using subsequently observed outcomes and evaluates temporal fidelity, thereby providing a closed-loop approximation of Digital Twin operation. The resulting predictive, sustainability, scenario, uncertainty, and temporal-fidelity information provides an integrated basis for interpretation and decision-oriented agricultural planning.

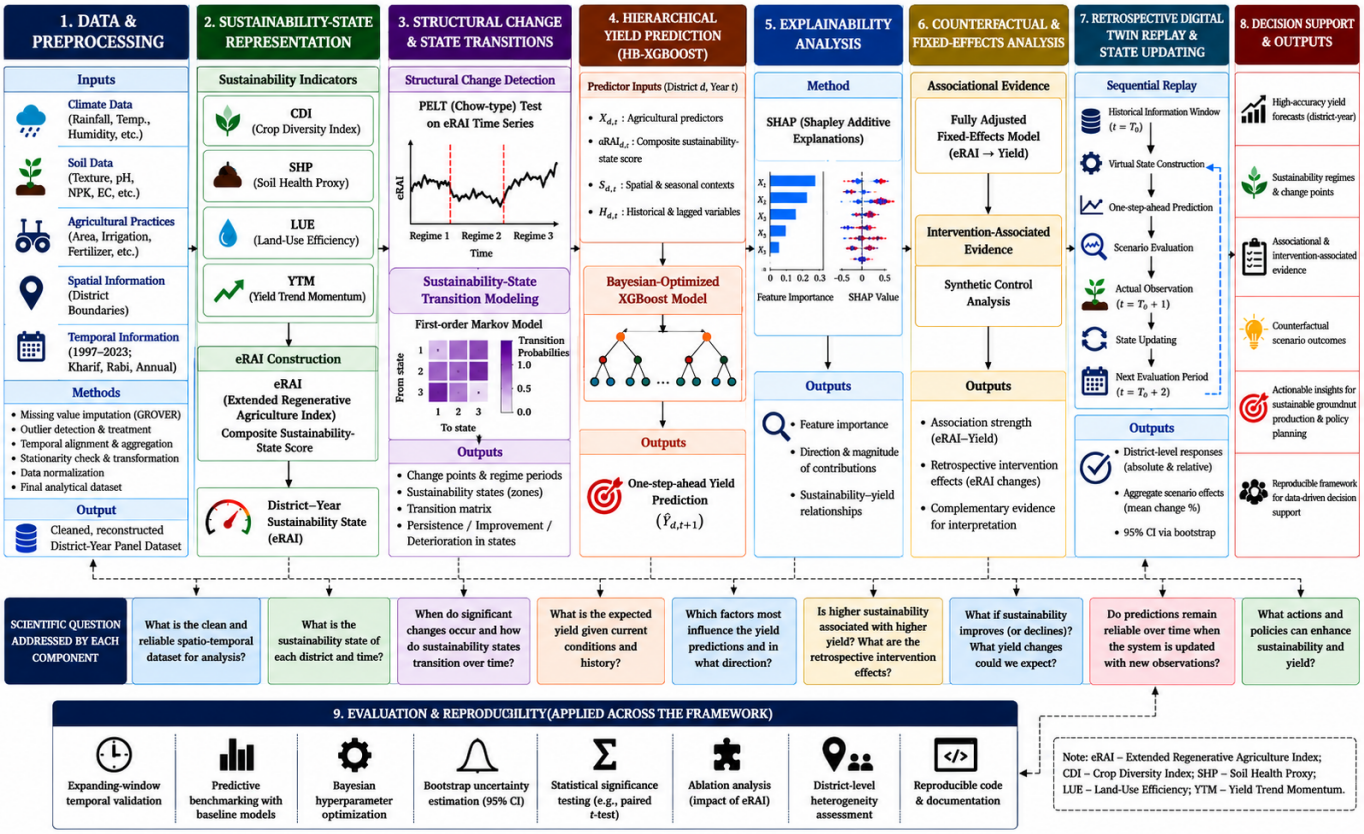

Figure 1. Overall workflow of the proposed TSAFI-DT framework.

The present implementation should be distinguished from a continuously synchronized cyber-physical Digital Twin. Because the experimental evaluation uses historical agricultural observations rather than live sensor streams, the physical system is represented retrospectively by temporally ordered district-level agricultural records. The corresponding virtual representation is updated sequentially using only information available up to each evaluation time. This design permits leakage-resistant forecasting and retrospective Digital Twin validation while providing an extensible foundation for future integration of Internet of Things (IoT), meteorological, and remote-sensing data.

2.1. Data Source and Study Representation

The experimental analysis uses the Area, Production, and Yield (APY) records published by the Directorate of Economics and Statistics, Department of Agriculture and Farmers Welfare, Government of India. Groundnut observations cover the period 1997–2023 and include district-, state-, year-, season-, crop-, cultivated-area-, production-, and yield-related information. Following data harmonization and eligibility screening, 412 districts are represented in the analytical dataset. For district d and year t, agricultural yield is represented as

$$Y_{d,t} = \frac{P_{d,t}}{A_{d,t}}, \tag{1}$$

where $P_{d,t}$ denotes agricultural production and $A_{d,t}$ denotes cultivated area. Records with invalid identifiers, non-positive area values, impossible yield values, or irreconcilable duplicates are removed before subsequent analysis.

District and state names are harmonized to account for spelling variation and administrative inconsistencies. Seasonal observations are standardized into consistent temporal categories. All transformations required for forecasting are performed within the corresponding historical training window to prevent future-information leakage.

2.2. Spatiotemporal Data Reconstruction

Historical agricultural records may contain missing observations and irregular temporal coverage. TSAFI-DT therefore applies a tiered reconstruction strategy that preserves temporal ordering and minimizes the use of future information.

Short internal gaps are reconstructed using temporally local information when sufficient neighbouring observations are available. For longer or structurally missing intervals, reconstruction uses historical district behaviour together with broader state- and season-level information available within the same training period. Variables required for sustainability-state construction are subsequently normalized using parameters estimated only from the historical training window. For a continuous variable x, min–max normalization is expressed as

$$x'_{d,t} = \frac{x_{d,t} - x_{\min}}{x_{\max} - x_{\min} + \epsilon}, \tag{2}$$

where $x_{\min}$ and $x_{\max}$ are estimated from the applicable training period and ϵ prevents division by zero. This temporal isolation ensures that reconstruction, scaling, feature engineering, and model optimization do not use information from future evaluation periods.

2.3. Structural-Change Detection

Agricultural productivity trajectories may exhibit regime changes arising from climatic variability, technological adoption, market conditions, structural agricultural changes, or policy periods. Candidate change points are identified using the Pruned Exact Linear Time (PELT) algorithm. For a district-level yield sequence, PELT estimates the change-point set $\mathcal{T}^*$ as

$$\mathcal{T}^* = \arg\min_{\mathcal{T}} \left[\sum_{k=0}^{K} C\left(Y_{\tau_k+1:\tau_{k+1}}\right) + \beta K \right]. \tag{3}$$

where $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$ denotes the set of detected change points, with $\tau_0 = 0$ and $\tau_{K+1} = T$ defining the boundaries of a yield sequence of length T. The term $C\left(Y_{\tau_k+1:\tau_{k+1}}\right)$ denotes the segment cost function for the observations from $\tau_k + 1$ through τ_{k+1}, K is the number of detected change points, and β is the penalty parameter controlling model

complexity. The PELT algorithm identifies the optimal segmentation by minimizing the total within-segment cost while penalizing excessive change points.

Candidate breaks are subsequently examined using a Chow-type structural stability assessment when sufficient observations are available on both sides of the detected point. PELT therefore identifies candidate temporal discontinuities, whereas the complementary stability assessment evaluates whether the corresponding change is statistically distinguishable. Detected break periods are interpreted as structural changes in historical trajectories. Temporal correspondence with agricultural programmes or policy periods is treated only as contextual evidence and is not interpreted as proof that a specific policy caused the detected change.

2.4. Extended Regenerative Agriculture Index

A sustainability state represents the multidimensional condition of an agricultural system at a given district and time. It should be distinguished from an individual sustainability indicator. In the proposed framework, the crop diversity index (CDI), soil health proxy (SHP), land-use efficiency (LUE), and yield trend momentum (YTM) are treated as individual sustainability-related indicators, while their joint representation characterizes the sustainability state of a district. The sustainability state is therefore represented as a multidimensional vector.

The term sustainability state refers to the collective multidimensional condition of an agricultural system for district d at time t, represented by the vector $S_{d,t}$. CDI, SHP, LUE, and YTM represent individual sustainability indicators or dimensions and are not separate sustainability states. The eRAI is subsequently derived as a composite representation of these dimensions through weighted aggregation. This formulation enables the sustainability condition of each district and year to be represented quantitatively and incorporated into the subsequent prediction and Digital Twin analyses. These indicators were selected to provide complementary views of agricultural sustainability, covering crop diversity, soil-related conditions, land-use efficiency, and temporal yield performance

$$\mathbf{S}_{i,t} = [\mathrm{CDI}_{i,t}, \mathrm{SHP}_{i,t}, \mathrm{LUE}_{i,t}, \mathrm{YTM}_{i,t}]. \quad (4)$$

Each component captures a distinct aspect of the agricultural system condition. CDI characterizes the degree of crop diversification and therefore provides an indication of diversity in the cropping system. SHP represents the soil-health-related condition captured by the constructed soil proxy and provides information on the productive condition of agricultural soils. LUE describes the efficiency with which available agricultural land is converted into crop production. YTM captures the temporal direction and momentum of yield performance, thereby distinguishing persistent yield improvement or deterioration from short-term fluctuations.

A central component of TSAFI-DT is the Extended Regenerative Agriculture Index (eRAI), which provides a composite sustainability-state representation using variables derivable from the available historical agricultural records. The index integrates the four complementary components, namely CDI, SHP, LUE, and YTM, into a composite measure. Thus, eRAI is treated as a composite sustainability indicator derived from the sustainability-state components rather than as an independent fifth dimension of the sustainability state. This distinction avoids double-counting the same underlying sustainability dimensions while allowing eRAI to provide a compact representation of the overall sustainability condition. Because the APY dataset does not contain direct laboratory measurements of soil physical or chemical properties, SHP is explicitly interpreted as a productivity–stability proxy rather than a direct measurement of soil health. The four components are compactly defined as

$$\mathrm{CDI}_{d,t} = 1 - \sum_{c=1}^{C} p_{c,d,t}^2, \qquad p_{c,d,t} = \frac{A_{c,d,t}}{\sum_{j=1}^{C} A_{j,d,t}}, \tag{5}$$

$$\mathrm{SHP}_{d,t} = 1 - \mathrm{Norm}\left(\frac{\sigma(Y_{d,t-w:t-1})}{\mu(Y_{d,t-w:t-1}) + \epsilon}\right), \tag{6}$$

$$\mathrm{LUE}_{d,t} = \mathrm{Norm}\left(\frac{\Delta Y_{d,t}}{|\Delta A_{d,t}| + \epsilon}\right), \tag{7}$$

$$\mathrm{YTM}_{d,t} = \mathrm{Norm}\left(\hat{\beta}_{1,d,t} + \lambda \hat{\beta}_{2,d,t}\right). \tag{8}$$

where C denotes the total number of crop categories considered in the crop-diversity calculation and is fixed for the analysis, while $p_{c,d,t}$ represents the cultivated-area proportion of crop category c in district d at time t. $p_{c,d,t}$ denotes the cultivated-area proportion of crop c in district d at time t, w denotes the trailing historical window, and $\mu(\cdot)$ and $\sigma(\cdot)$ denote the historical mean and standard deviation, respectively. The quantities $\Delta Y_{d,t}$ and $\Delta A_{d,t}$ represent temporal changes in yield and cultivated area, respectively, while $\hat{\beta}_{1,d,t}$ and $\hat{\beta}_{2,d,t}$ characterize the local slope and curvature of the historical yield trajectory. The operator $\mathrm{Norm}(\cdot)$ maps each component to a common scale using parameters derived from the applicable historical training data.

The selection of these indicators is motivated by their complementary representation of agricultural sustainability, encompassing crop diversity, soil-related productivity conditions, land-use efficiency, and temporal yield performance [34–37]. The normalized components are integrated to obtain the Extended Regenerative Agriculture Index:

$$\mathrm{eRAI}_{d,t} = \omega_1 \mathrm{CDI}_{d,t} + \omega_2 \mathrm{SHP}_{d,t} + \omega_3 \mathrm{LUE}_{d,t} + \omega_4 \mathrm{YTM}_{d,t}, \qquad \sum_{k=1}^{4} \omega_k = 1, \tag{9}$$

where ω_k denotes the contribution weight assigned to the corresponding sustainability dimension. The resulting eRAI provides a compact composite representation of the multi-dimensional sustainability condition while retaining the underlying component indicators for interpretation and perturbation analysis.

To assess the robustness of the predefined eRAI weighting scheme, a sensitivity analysis was conducted using alternative weight configurations while maintaining the constraint $\sum_{k=1}^{4} \omega_k = 1$. The alternative configurations were designed to modify the relative contribution of CDI, SHP, LUE, and YTM without changing the underlying component definitions or normalization procedure. The resulting eRAI variants were subsequently evaluated for consistency in sustainability-state classification and predictive performance.

In the predictive stage, the sustainability-related indicators and their derived representations are incorporated into the model feature space together with the other explanatory variables. In the Digital Twin stage, selected sustainability-state components can be virtually perturbed while holding the remaining model structure fixed. This allows the sensitivity of predicted yield to controlled changes in specific sustainability dimensions to be examined. The resulting scenario outputs are interpreted as model-based sensitivity estimates for comparative scenario exploration rather than as causal intervention effects or experimentally verified yield gains.

2.5. Sustainability-State Transition Modelling

District-level eRAI trajectories are used to characterize temporal sustainability dynamics. District-year observations are assigned to ordered sustainability states according to the distribution of eRAI within the applicable analytical configuration. Temporal persistence

and movement between sustainability states are summarized using a first-order Markov transition model:

$$P_{ij} = P(S_{t+1} = j \mid S_t = i) = \frac{N_{ij}}{\sum_k N_{ik}}, \tag{10}$$

where S_t represents the sustainability state at time t, N_{ij} is the number of observed transitions from state i to state j, and P_{ij} is the corresponding transition probability. The transition matrix is used descriptively to identify persistence, improvement, and deterioration in sustainability-state trajectories. These probabilities characterize historical state dynamics and are not interpreted as causal effects of interventions.

2.6. Hierarchical Yield-Prediction Model

The predictive layer employs a hierarchical Bayesian-optimized Extreme Gradient Boosting model (HB-XGBoost). Recent agricultural prediction research has demonstrated the effectiveness of XGBoost-based frameworks for robust crop-yield prediction under uncertainty, including feature optimization and uncertainty-aware modelling [38]. The model combines historical productivity, cultivated-area information, spatial context, seasonal information, sustainability-state variables, and lagged temporal features. For district d at time t, one-step-ahead yield prediction is expressed as

$$\widehat{Y}_{d,t+1} = F_{\theta^*}(\mathbf{X}_{d,t}, eRAI_{d,t}, S_{d,t}, \mathbf{H}_{d,t}), \tag{11}$$

where $\mathbf{X}_{d,t}$ contains the available agricultural predictors, $eRAI_{d,t}$ is the composite sustainability-state score derived from the underlying sustainability indicators, $S_{d,t}$ represents spatial and seasonal contextual information, $\mathbf{H}_{d,t}$ contains historical and lagged variables, and F_{θ^*} denotes the optimized XGBoost predictor. Separate contextual representations are considered for Kharif, Rabi, and annual configurations to account for season-specific differences in predictor importance. Feature selection and hyperparameter optimization are performed using only the training portion of each temporal evaluation cycle.

Bayesian optimization is used to identify an effective hyperparameter configuration for the XGBoost predictor. The optimized parameters include the number of estimators, learning rate, maximum tree depth, subsampling ratio, column-sampling ratio, and regularization parameters. Optimization is nested within the historical training data so that information from the evaluation period does not influence parameter selection.

2.7. Explainability Analysis

SHapley Additive exPlanations (SHAPs) are used to examine the contribution of individual predictors to HB-XGBoost predictions. For prediction $f(\mathbf{x})$, the additive explanation is represented as

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{M} \phi_j, \tag{12}$$

where j denotes the index of an individual sustainability-state component, and M denotes the total number of sustainability-state components included in the formulation. Thus, the summation from $j = 1$ to M accounts for all sustainability-state components considered in the transition analysis. ϕ_0 denotes the expected model output and ϕ_j represents the SHAP contribution of feature j. The analysis is used to identify dominant predictors and examine nonlinear relationships among sustainability-state variables, previous yield, temporal momentum, and other contextual features. SHAP values are interpreted as model-explanation quantities and not as estimates of causal effects.

2.8. Counterfactual and Fixed-Effects Analysis

Counterfactual and panel analyses are conducted separately from predictive model evaluation. These analyses examine intervention-associated trajectory differences and adjusted relationships between sustainability states and agricultural yield. They are not used to generate the predictive RMSE reported for HB-XGBoost. For the counterfactual analysis, a synthetic comparison trajectory is constructed as a weighted combination of eligible comparison units:

$$Y_{d,t}^{SC} = \sum_{j \in \mathcal{C}_d} w_j Y_{j,t}, \qquad w_j \geq 0, \qquad \sum_j w_j = 1, \tag{13}$$

The donor pool C_d for district d comprises the eligible comparison districts used to construct the synthetic control. Candidate districts are selected from the same study population and observation period and must have sufficient pre-intervention observations to support the counterfactual estimation. The treated district d itself is excluded from its donor pool. Districts with inadequate pre-period data or insufficient overlap with the treated district in the relevant pre-intervention characteristics are also excluded. The resulting set C_d provides the comparison units from which the synthetic control for district d is constructed. w_j represents the non-negative weight assigned to comparison unit j. The weights are selected to minimize pre-transition differences between the evaluated trajectory and its synthetic counterpart. The post-transition difference is summarized as

$$\Delta_{d,t}^{SC} = Y_{d,t} - Y_{d,t}^{SC}. \tag{14}$$

The resulting difference is interpreted as intervention-associated counterfactual evidence subject to the adequacy of pre-transition fit and the validity of the comparison structure. It is not interpreted as unconditional proof that a change in eRAI caused the observed yield difference. Complementary panel analysis is performed using nested district- and year-fixed-effects specifications:

$$Y_{d,t} = \alpha_d + \gamma_t + \beta H_{d,t} + \boldsymbol{\delta}^\top \mathbf{C}_{d,t} + \varepsilon_{d,t}, \tag{15}$$

where α_d denotes district fixed effects, γ_t denotes year fixed effects, $H_{d,t}$ represents the high-sustainability-state indicator, $\mathbf{C}_{d,t}$ contains the controls included in the corresponding nested model, and $\varepsilon_{d,t}$ is the error term. Four nested model configurations are evaluated. Model 1 estimates the baseline relationship between high sustainability state and yield. Model 2 introduces temporal fixed effects. Model 3 incorporates district and temporal controls to account for persistent spatial heterogeneity. Model 4 represents the fully adjusted specification by additionally including lagged yield and lagged cultivated area. Because high eRAI is derived partly from historical outcome-related information and is not an independently assigned treatment, the coefficient β is interpreted as an adjusted association rather than a conventional Difference-in-Differences treatment effect. This distinction avoids unsupported causal attribution.

2.9. Digital Twin Scenario Simulation

The Digital Twin simulation layer evaluates the sensitivity of predicted yield to predefined hypothetical modifications of sustainability-related virtual state variables. The trained predictive model is retained unchanged, while one or more state variables are perturbed. For a baseline state vector $\mathbf{z}_{d,t}$ and a scenario-specific perturbation $\Delta\mathbf{z}^{(s)}$, the scenario prediction is

$$\widehat{Y}_{d,t+1}^{(s)} = F_{\theta^*}\left(\mathbf{z}_{d,t} + \Delta\mathbf{z}^{(s)}\right). \tag{16}$$

Three configurations are evaluated: a baseline with no perturbation, a moderate scenario involving a 10% increase in SHP, and a high scenario involving simultaneous 20% increases in SHP and CDI. The relative model-based change is calculated as

$$\Delta Y^{(s)}(\%) = 100\left(\frac{\overline{\widehat{Y}}^{(s)} - \overline{\widehat{Y}}^{(0)}}{\overline{\widehat{Y}}^{(0)}}\right), \tag{17}$$

where $\overline{\widehat{Y}}^{(0)}$ is the mean baseline prediction and $\overline{\widehat{Y}}^{(s)}$ is the mean prediction under scenario s. These scenario outputs are interpreted strictly as model-based counterfactual sensitivity estimates. They quantify the response of the trained model to predefined changes in virtual sustainability states and do not represent experimentally observed gains or guaranteed causal effects of real-world interventions.

The Digital Twin perturbation analysis is performed separately for each district-year observation using its corresponding sustainability-state representation. For district d under scenario s, the perturbation response is defined as the difference between the perturbed prediction and the corresponding unperturbed baseline prediction:

$$\Delta Y_d^{(s)} = \hat{Y}_d^{(s)} - \hat{Y}_d^{(0)}. \tag{18}$$

The corresponding relative perturbation response is calculated as

$$\Delta Y_d^{(s)}(\%) = \frac{\hat{Y}_d^{(s)} - \hat{Y}_d^{(0)}}{\hat{Y}_d^{(0)}} \times 100. \tag{19}$$

Thus, the perturbation is evaluated relative to the baseline prediction of the same district rather than assuming a common baseline across all districts. The reported scenario-level effects are subsequently obtained by aggregating the district-level responses. Consequently, the reported percentage changes represent an aggregate model-based response across the evaluated districts, while individual district responses may vary according to their respective sustainability states and other explanatory characteristics.

To quantify the uncertainty associated with the estimated perturbation responses, baseline and scenario predictions are evaluated as paired predictions for the same district-year observations. For each observation, the difference between the perturbed and baseline prediction is calculated, and the corresponding relative change is obtained. The distribution of these paired perturbation effects is used to estimate the uncertainty of the aggregate scenario response. A 95% confidence interval is reported for each scenario, together with the mean and standard deviation of the district-level responses. This procedure allows the magnitude and variability of the scenario response to be assessed independently from the overall predictive RMSE. The perturbation results are interpreted as model-based sensitivity estimates rather than experimentally verified causal effects.

2.10. Retrospective Digital Twin State Updating

The Digital Twin feedback mechanism is implemented through retrospective sequential state updating. At each historical evaluation point, the virtual state is reconstructed using only observations available up to that time. A one-step-ahead prediction is generated and subsequently compared with the observed outcome when it becomes available in the historical sequence. The update cycle can be summarized as

$$\mathcal{V}_{t+1} = \mathcal{U}\left(\mathcal{V}_t, Y_{t+1}^{obs}, \mathbf{X}_{t+1}\right), \tag{20}$$

where $\mathcal{V}_t$ denotes the virtual state at time t, Y_{t+1}^{obs} is the subsequently observed yield, $\mathbf{X}_{t+1}$ contains newly available historical information, and $\mathcal{U}(\cdot)$ denotes the retrospective state-update operation. This sequential replay approximates the computational feedback cycle of a Digital Twin while maintaining a clear distinction from real-time cyber-physical synchronization. Temporal fidelity is evaluated by comparing one-step-ahead predictions with subsequently observed historical yields.

2.11. Experimental Setup and Reproducibility

The experimental design follows an expanding-window temporal evaluation protocol rather than random cross-validation. For each evaluation cycle, all observations available before the target period constitute the training set, and the subsequent period is reserved for out-of-sample testing. Data reconstruction, normalization, sustainability-state construction, feature processing, feature selection, and Bayesian hyperparameter optimization are performed within each historical training window.

The expanding-window design separates historical model development from subsequent out-of-sample evaluation. At each evaluation cycle t, all observations available up to time t are used as the training set, while observations from the immediately subsequent period $t + 1$ are retained exclusively for out-of-sample testing. Bayesian hyperparameter optimization is performed only within the historical training window and is therefore not exposed to observations from the subsequent test period. Similarly, data reconstruction, normalization, sustainability-state construction, feature processing, and feature selection are performed independently within each historical training window. The selected model is then applied without further fitting to the subsequent unseen period, and MAE, RMSE, and R^2 are calculated using only these out-of-sample predictions. After evaluation, the newly available observations are incorporated into the expanded historical window for the next evaluation cycle. Thus, no future observations are used during preprocessing, feature selection, or hyperparameter optimization, thereby maintaining chronological integrity and reducing the risk of information leakage.

No conventional random train–test split or random k-fold cross-validation was used, because such procedures could mix observations from different time periods and potentially introduce temporal information leakage. Instead, the expanding-window temporal protocol provides repeated out-of-sample evaluation while preserving the chronological structure of the agricultural data.

The proposed HB-XGBoost predictor is compared with three representative baselines: Linear Regression as a statistical baseline, conventional XGBoost as a machine-learning baseline, and GA+XGBoost as an optimization-based baseline. All models are evaluated under equivalent temporal partitions to ensure a fair comparison. Predictive performance is evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2):

$$\begin{aligned} MAE &= \frac{1}{N}\sum_{i=1}^{N}|Y_i - \widehat{Y}_i|, \\ RMSE &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2}, \\ R^2 &= 1 - \frac{\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}. \end{aligned} \tag{21}$$

Confidence intervals across temporal evaluation cycles are used to assess performance stability. Paired parametric and non-parametric statistical tests are used, where appropriate,

to evaluate whether the performance differences between HB-XGBoost and the baseline predictors are statistically significant at the adopted significance level.

Temporal Training, Validation, and Testing

One-step-ahead yield forecasting requires preservation of the chronological structure of the observations; therefore, the data were not randomly divided into training and test sets, and conventional random k-fold cross-validation was not employed. Instead, an expanding-window temporal evaluation strategy was adopted to prevent information from future periods from entering model development. At each evaluation cycle t, all district-year observations available before the target period were used as the historical training set, while the immediately subsequent period was retained as the out-of-sample test set. After evaluation, the newly observed period was incorporated into the training window, and the procedure was repeated for the next forecast period. Thus, the training window expands chronologically across evaluation cycles, whereas each test period remains unseen during model development.

No separate randomly sampled validation set was created. For hyperparameter selection, the historical training window was internally used by the Bayesian optimization procedure, while the subsequent chronological period was never used for parameter selection and served exclusively as the out-of-sample evaluation period. Instead, model selection and hyperparameter optimization were performed within the historical training window, using only information available before the corresponding test period. Bayesian hyperparameter optimization was therefore nested within each training window, and the selected configuration was subsequently evaluated on the next unseen period. The same temporal training and testing partitions were applied to HB-XGBoost, Linear Regression, conventional XGBoost, and GA+XGBoost to ensure a fair comparison. All preprocessing operations, including normalization, feature selection, and sustainability-state construction, were fitted using the corresponding historical training data before being applied to the subsequent test period.

To provide stronger temporal benchmarks, Lag-Yield and Lag-Yield + Area models were additionally evaluated using the same expanding-window temporal validation protocol. A separate spatial forecasting baseline was not introduced because the proposed pooled framework operates at the district level and incorporates district-specific effects while maintaining common temporal partitions across districts. This design prioritizes temporal generalization under a consistent pooled forecasting setting.

2.12. Predictive Ablation Analysis

Ablation analysis is restricted to components that directly participate in the yield-prediction pipeline. The full HB-XGBoost configuration is compared with reduced variants in which selected predictive information is removed, and the resulting change in RMSE is evaluated. The relative degradation in predictive performance is calculated as

$$\Delta RMSE(\%) = 100\left(\frac{RMSE_{\text{ablated}} - RMSE_{\text{full}}}{RMSE_{\text{full}}}\right). \quad (22)$$

Counterfactual and fixed-effects analytical modules are not treated as predictive ablation components because they operate downstream of yield forecasting and do not directly determine predictive RMSE. This separation prevents forecasting improvements from being incorrectly attributed to analytical modules that are not involved in generating the yield predictions.

Overall, the Materials and Methods design maintains a strict separation among three analytical objectives: HB-XGBoost provides out-of-sample yield forecasting; counterfactual and fixed-effects analyses provide intervention-associated and adjusted associational evi-

dence; and the Digital Twin simulation layer provides model-based "what-if" sensitivity analysis. This separation ensures that predictive performance, adjusted associations, and simulated scenario responses are interpreted according to the evidence provided by their respective analytical components.

3. Results and Discussion

This section evaluates the TSAFI-DT framework in terms of data reconstruction, structural-change detection, sustainability-state characterization, predictive performance, counterfactual analysis, spatial sustainability patterns, scenario simulation, retrospective temporal fidelity, and component-level ablation. The results are interpreted according to the role of each framework component. Predictive performance is evaluated using out-of-sample forecasting metrics, whereas sustainability transitions and scenario analyses are interpreted as descriptive or model-based evidence unless an independently defined intervention supports causal interpretation.

3.1. Data Reconstruction and Structural-Change Analysis

The APY groundnut panel exhibits substantial missingness arising from incomplete and temporally inconsistent district-level reporting. The tiered reconstruction strategy was therefore evaluated using controlled masking of originally observed records, allowing reconstruction error to be measured against known values.

Table 1 summarizes the reconstruction results. Approximately 63% of the observations were available in complete form, while the remaining records required reconstruction at different levels of missingness. Temporal interpolation achieved the lowest reconstruction error for isolated gaps, with an RMSE of 0.018 t/ha. Spatial-neighbour reconstruction obtained an RMSE of 0.027 t/ha, whereas matrix factorization produced an RMSE of 0.031 t/ha for extended missing blocks. These results indicate that reconstruction performance depends on the structure of missingness and support the use of a tier-specific strategy rather than a single imputation mechanism.

Table 1. Tiered reconstruction summary across the APY groundnut dataset (1997–2023).

Tier	Strategy	Districts	District-Years	RMSE (t/ha)
0	Complete data	260	3380	—
1	Temporal interpolation	87	521	0.018
2	Spatial-neighbour imputation	45	312	0.027
3	Matrix factorization	20	198	0.031

Structural-change analysis further revealed substantial temporal heterogeneity in district-level yield trajectories. Table 2 summarizes the distribution of detected break periods. The largest concentration occurred during 2012–2014, with 34 districts exhibiting structural changes, while additional clusters were observed during 2003–2005, 2007–2009, and 2018–2020.

Table 2. Distribution of structural-break periods across districts.

Break-Year Range	Districts	Concurrent Policy Context	Yield Change
2003–2005	11	National Oilseeds Mission	+0.07
2007–2009	18	RKVY	+0.11
2012–2014	34	NFSA + MSP revision	+0.18
2018–2020	22	PM-AASHA	+0.14
No detected break	327	—	—

The temporal coincidence between several detected breaks and major agricultural policy periods provides useful contextual evidence; however, temporal coincidence alone does not establish policy causality. More generally, the detected structural changes demonstrate that groundnut yield trajectories are non-stationary across districts. This finding supports the use of temporally adaptive forecasting and state-transition analysis within TSAFI-DT.

3.2. Extended Regenerative Agriculture Index and Sustainability Dynamics

The Extended Regenerative Agriculture Index (eRAI) integrates crop diversity index (CDI), soil health proxy (SHP), land-use efficiency (LUE), and yield trend momentum (YTM) into a unified sustainability-state representation. Table 3 reports the descriptive statistics and associations of each component with next-year yield.

Among the individual components, YTM exhibited the strongest correlation with next-year yield ($r = 0.63$), followed by SHP ($r = 0.55$), LUE ($r = 0.51$), and CDI ($r = 0.48$). The integrated eRAI achieved a higher correlation of $r = 0.74$, indicating that the combined representation captures complementary information that is not fully represented by any individual component.

Table 3. eRAI component statistics across districts.

Component	Weight	Mean	Std. Dev.	Corr. (Yield$_{t+1}$)
CDI	0.35	0.68	0.12	0.48
SHP	0.28	0.71	0.09	0.55
LUE	0.22	0.43	0.17	0.51
YTM	0.15	0.09	0.06	0.63
eRAI	—	0.59	0.11	0.74

To assess the robustness of the predefined eRAI weighting scheme, alternative weight configurations were evaluated while maintaining the constraint $\sum_{k=1}^{4} \omega_k = 1$. The resulting eRAI variants were compared with the baseline formulation in terms of predictive performance.

The sensitivity analysis indicates that predictive performance remains broadly stable under the alternative eRAI weighting configurations and is shown in Table 4. Relative to the baseline weighting scheme, RMSE varied from 0.171 to 0.183 and R^2 from 0.92 to 0.90 across the tested configurations. These modest variations suggest that the principal predictive conclusions are not driven solely by the specific a priori eRAI weights.

Table 4. Sensitivity of eRAI to alternative weighting configurations.

Configuration	ω_{CDI}	ω_{SHP}	ω_{LUE}	ω_{YTM}	RMSE	R^2
Baseline	0.35	0.28	0.22	0.15	0.171	0.92
Alternative 1	0.25	0.25	0.25	0.25	0.183	0.90
Alternative 2	0.25	0.25	0.20	0.30	0.179	0.91

Figure 2 illustrates representative temporal eRAI trajectories. The selected districts exhibit distinct patterns of improvement, decline, and relative stability, demonstrating that sustainability states evolve heterogeneously across locations and time. The PELT-detected change points further identify periods in which the trajectory of the sustainability representation changes substantially.

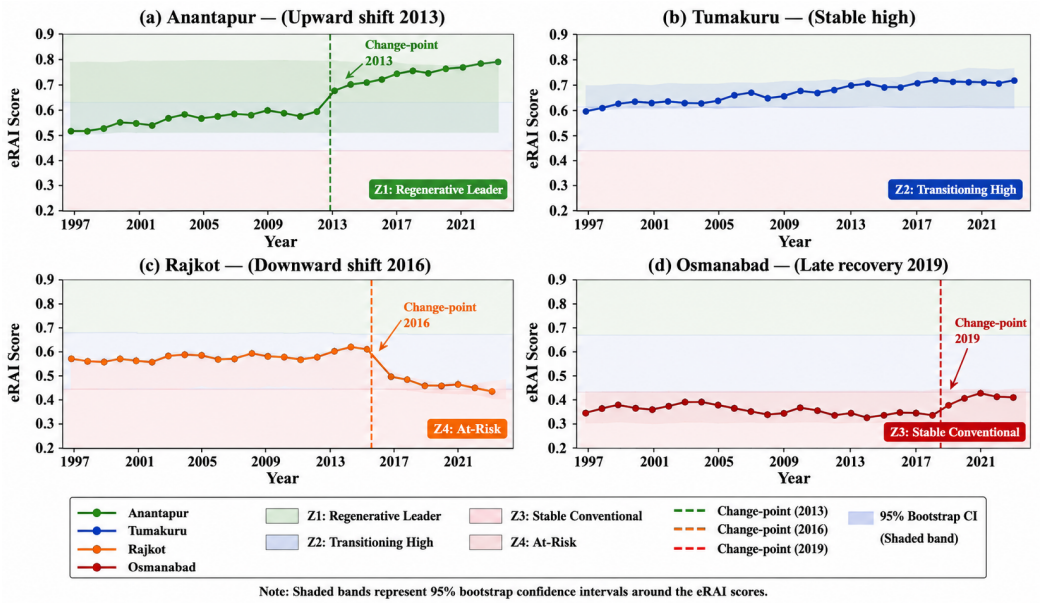

Figure 2. Temporal trajectories of eRAI scores for representative districts with PELT-detected change points.

The Markov transition probabilities reported in Table 5 provide additional evidence of persistence in sustainability states. Districts in the high state remained in that state with probability 0.79, while districts in the low state remained low with probability 0.81. The medium state exhibited lower persistence (0.62) and greater movement toward neighbouring states.

Table 5. Markov sustainability-state transition matrix.

From/To	High	Medium	Low
High	0.79	0.17	0.04
Medium	0.21	0.62	0.17
Low	0.06	0.13	0.81

The high persistence probabilities suggest substantial temporal inertia, particularly at the two extremes of the sustainability distribution. These transition probabilities should be interpreted as descriptive state dynamics rather than causal effects. Nevertheless, they identify districts where sustainability conditions appear persistent and may therefore warrant closer monitoring or targeted scenario assessment.

3.3. Hierarchical Yield-Prediction Performance

The predictive contribution of sustainability, temporal, spatial, and season-specific information was evaluated using the hierarchical Bayesian-optimized XGBoost (HB-XGBoost) model. Table 6 presents the highest-ranked predictors across Kharif, Rabi, and annual configurations. eRAI and previous-year yield consistently appear among the dominant predictors, while YTM also contributes across all three configurations. The remaining features vary by season, indicating that the relative importance of spatial and agronomic information is not uniform across seasonal contexts.

Table 6. Selected feature subsets across seasonal prediction configurations.

Rank	Kharif	Rabi	Annual
1	eRAI Score	eRAI Score	eRAI Score
2	Previous Yield	Previous Yield	Previous Yield
3	District Effect	ΔY_{season}	District Effect
4	YTM	YTM	YTM
5	Area	State eRAI	CDI

Table 7 compares HB-XGBoost with Linear Regression, conventional XGBoost, and GA+XGBoost under the same temporal evaluation framework. HB-XGBoost achieved the lowest MAE (0.128 t/ha) and RMSE (0.171 t/ha), together with the highest R^2 (0.92). Relative to GA+XGBoost, the RMSE decreased from 0.204 to 0.171 t/ha, corresponding to an improvement of approximately 16.2%.

Table 7. Predictive performance comparison of the evaluated models.

Model	MAE	RMSE	R^2	Description
Linear Regression	0.221	0.291	0.78	Statistical baseline
XGBoost	0.184	0.239	0.85	ML baseline
GA+XGBoost	0.149	0.204	0.89	Optimization baseline
HB-XGBoost	0.128	0.171	0.92	Proposed predictor

The confidence-interval analysis in Table 8 shows that HB-XGBoost maintained the lowest mean RMSE, with a 95% confidence interval of 0.164–0.178. Its interval does not overlap with those of the evaluated baselines, indicating consistent performance differences across the temporal evaluation cycles.

Table 8. Confidence intervals for predictive RMSE.

Model	Mean RMSE	95% CI Lower	95% CI Upper
Linear Regression	0.291	0.278	0.304
XGBoost	0.239	0.228	0.250
GA+XGBoost	0.204	0.195	0.213
HB-XGBoost	0.171	0.164	0.178

The paired significance tests reported in Table 9 further support these differences. HB-XGBoost significantly outperformed Linear Regression, XGBoost, and GA+XGBoost under both the paired *t*-test and Wilcoxon signed-rank test ($p < 0.05$ for all comparisons).

Table 9. Statistical significance analysis of predictive performance.

Comparison	*t*-Test (*p*-Value)	Wilcoxon (*p*-Value)	Significant
HB-XGBoost vs. Linear Regression	<0.001	<0.001	Yes
HB-XGBoost vs. XGBoost	0.003	0.005	Yes
HB-XGBoost vs. GA+XGBoost	0.012	0.018	Yes

The SHAP analysis in Figure 3 provides an interpretable view of the predictive model. eRAI exhibits the largest overall contribution, followed by previous-year yield and district-level effects, while YTM also contributes substantially to model output. The distribution of SHAP values indicates that the model uses sustainability and historical productivity information jointly rather than relying on a single predictor.

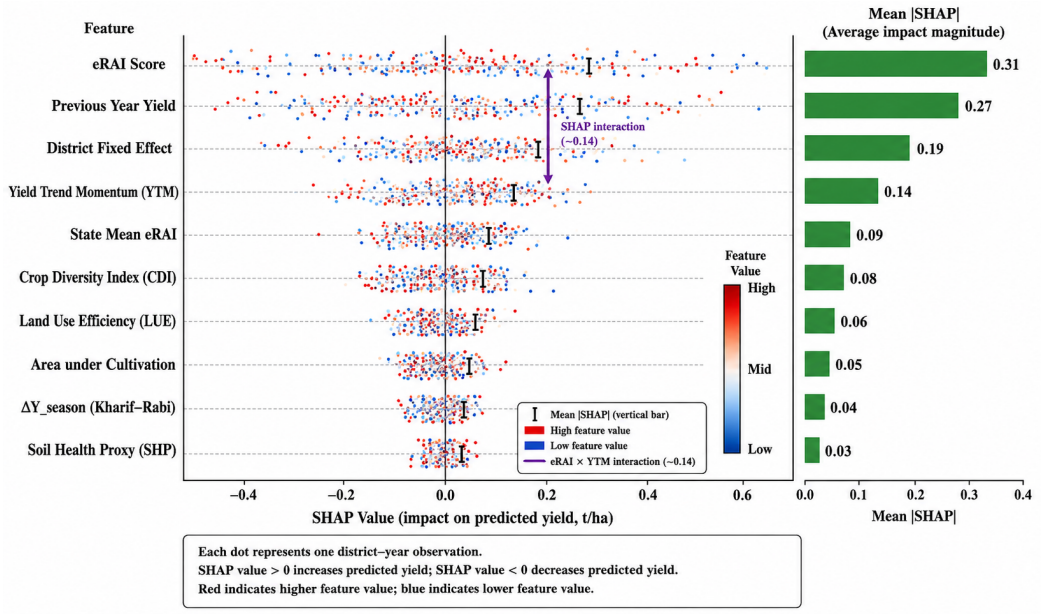

Figure 3. SHAP-based feature importance and contribution patterns for the pooled HB-XGBoost model.

Overall, the predictive results demonstrate that combining historical yield, spatial context, temporal dynamics, and the integrated sustainability-state representation improves one-step-ahead forecasting relative to the evaluated baseline models.

3.4. Counterfactual and Intervention-Associated Analysis

The counterfactual layer was examined separately from predictive evaluation to avoid interpreting outcome-derived sustainability indicators as causal treatments. Figure 4 compares the observed trajectory of the evaluated transition group with its synthetic counterfactual. Following the transition point, the observed and synthetic trajectories diverge, with an estimated average post-transition difference of +0.19 t/ha and a 95% confidence interval of +0.13 to +0.25 t/ha.

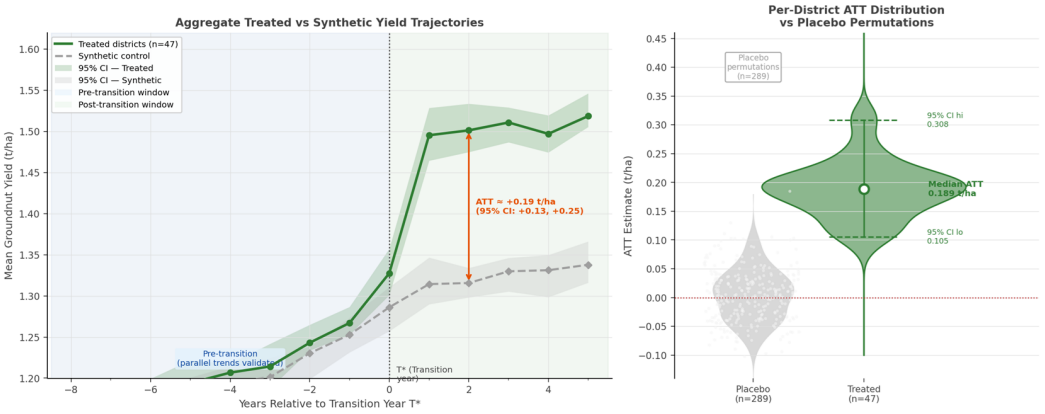

Figure 4. Observed and synthetic counterfactual yield trajectories around the evaluated sustainability transition.

Complementary fixed-effects estimates are reported in Table 10. Model 1 evaluates the baseline association between the high-eRAI state and yield. Model 2 introduces district fixed effects to account for time-invariant district-level heterogeneity, while Model 3 additionally incorporates year fixed effects to control for common temporal shocks. Model 4 provides the fully adjusted specification by including lagged yield and lagged cultivated area.

The estimated coefficient for the high-eRAI state remains positive across all four specifications, decreasing from 0.24 in Model 1 to 0.17 in Model 4 as additional controls are introduced. The fully adjusted model achieves the highest explanatory power ($R^2 = 0.74$). These results indicate a robust positive association between higher sustainability states and groundnut yield after accounting for observed spatial and temporal heterogeneity.

Table 10. Nested fixed-effects estimates for the association between sustainability state and groundnut yield.

Variable	Model 1	Model 2	Model 3	Model 4
High eRAI	0.24 ***	0.21 ***	0.18 ***	0.17 ***
Lag Yield	—	—	—	0.43 ***
Lag Area	—	—	—	0.008 *
District FE	No	Yes	Yes	Yes
Year FE	No	No	Yes	Yes
Additional Controls	No	No	No	Yes
Observations	8624	8624	8624	8177
R^2	0.11	0.24	0.61	0.74

Notes: Statistical significance: * $p < 0.05$, *** $p < 0.001$.

These results indicate a statistically significant association between higher sustainability states and improved yield outcomes. However, because eRAI contains components derived partly from historical yield behaviour, the high-eRAI classification itself is not treated as an independently assigned intervention. Accordingly, the estimates reported here should be interpreted as intervention-associated or quasi-experimental evidence unless the transition can be linked to an independently documented treatment with a clearly established timing and valid identifying assumptions. This interpretation is consistent with the causal safeguards defined in the methodology.

3.5. Pan-India Sustainability-State Analysis

The district-level sustainability classification reveals substantial spatial heterogeneity across the 412 evaluated districts. Table 11 summarizes the five sustainability zones. A clear gradient is observed from Z1 to Z5: mean eRAI decreases from 0.79 to 0.31, mean yield declines from 1.68 to 1.02 t/ha, and mean YTM shifts from +0.12 to −0.09.

Table 11. Pan-India sustainability-zone statistics.

Zone	Districts	Mean eRAI	Mean Yield (t/ha)	Mean YTM
Z1	52	0.79	1.68	+0.12
Z2	81	0.67	1.51	+0.07
Z3	112	0.54	1.38	+0.01
Z4	98	0.41	1.19	−0.04
Z5	69	0.31	1.02	−0.09

The Z4 and Z5 categories contain 167 districts, corresponding to approximately 40.5% of the evaluated districts. Their lower eRAI values and negative mean trend momentum indicate comparatively less favourable sustainability trajectories. In contrast, Z1 and Z2 combine higher sustainability-state values with stronger average productivity and positive yield momentum.

Figure 5 provides a spatial representation of these sustainability patterns and illustrates the heterogeneity of district-level agricultural states across the evaluated regions.

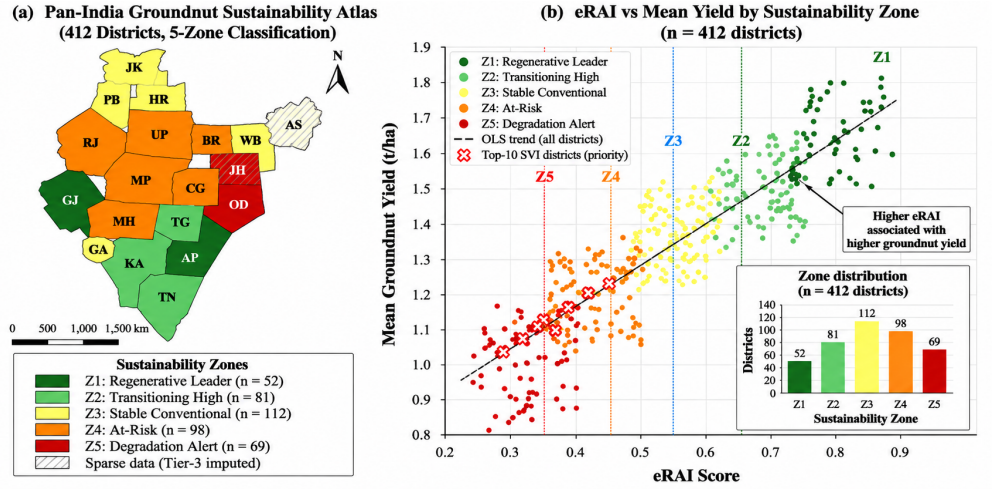

Figure 5. Pan-India representation of groundnut sustainability zones derived from district-level eRAI and productivity trajectories.

The zone analysis should be interpreted as a comparative sustainability classification rather than a direct measurement of ecological condition, particularly because SHP is a productivity–stability proxy rather than a direct soil measurement. Nevertheless, the classification provides a useful mechanism for identifying districts with persistently weak or improving sustainability trajectories for subsequent monitoring and scenario analysis.

3.6. Result of Digital Twin Scenario Simulation

The scenario-simulation layer evaluates the response of the trained Digital Twin predictor to hypothetical modifications of sustainability-related state variables. Three configurations were examined: the observed baseline, a moderate scenario with a 10% increase in SHP, and a high scenario with simultaneous 20% improvements in SHP and CDI.

The scenario variables differ in their practical actionability. SHP-related dimensions can be influenced indirectly through farm-level management practices that support crop diversification, resource-use efficiency, and production stability, whereas CDI-related changes may require longer-term agronomic and institutional interventions. In contrast, YTM is primarily a historical productivity-derived indicator and should not be interpreted as a directly controllable intervention variable. Therefore, the simulated perturbations should be regarded as model-based "what-if" sensitivity estimates rather than directly implementable interventions or guaranteed yield improvements.

Table 12 shows that the baseline mean predicted yield was 1.42 t/ha, with a 95% bootstrap confidence interval of 1.39–1.45 t/ha. Under the moderate scenario, the mean predicted yield increased to 1.52 t/ha, corresponding to a 6.8% increase, with a 95% confidence interval of 1.48–1.56 t/ha. Under the combined high scenario, the mean predicted yield reached 1.60 t/ha, representing a 12.4% increase relative to the baseline, with a 95% confidence interval of 1.55–1.64 t/ha. The non-overlapping bootstrap confidence intervals provide supporting evidence that the scenario-level mean predicted yields are distinguishable, while the reported perturbation responses remain model-based sensitivity estimates rather than causal forecasts. These simulated improvements should not be interpreted as causal treatment effects or as direct estimates of policy impact. The scenarios represent counterfactual perturbations within the fitted forecasting model and do not account for implementation costs, farmer adoption constraints, climatic variability, institutional conditions, or possible interactions among interventions. Field-level validation would therefore be required before using such scenario estimates to inform operational agricultural policy or management decisions.

Uncertainty interpretation. The bootstrap confidence intervals quantify the variability of the estimated scenario-level predicted yields resulting from resampling the model-training data. They are therefore used to characterize uncertainty in the scenario estimates rather than to provide formal hypothesis tests of the paired differences between baseline and perturbed predictions. The perturbation effects are interpreted by comparing paired district-year predictions under the baseline and corresponding perturbed states, while the out-of-sample RMSE is reported separately as a measure of predictive error. Accordingly, the perturbation results should be interpreted as model-based sensitivity estimates rather than as statistically verified causal yield gains.

Table 12. Digital Twin scenario-simulation results under sustainability-state perturbations with 95% bootstrap confidence intervals.

Scenario	Simulated Change	Mean Yield (t/ha)	95% CI (t/ha)	Yield Change (%)
Baseline	No perturbation	1.42	1.39–1.45	–
Moderate	+10% SHP	1.52	1.48–1.56	+6.8%
High	+20% SHP and CDI	1.60	1.55–1.64	+12.4%

The scenario results represent deterministic forward-pass sensitivity estimates rather than out-of-sample prediction intervals. The baseline and perturbed sustainability states are passed through the trained HB-XGBoost model while applying the specified virtual perturbations, whereas the out-of-sample RMSE reflects the prediction error of the model on previously unseen observations. The RMSE therefore characterizes predictive performance on unseen data, while the scenario differences quantify the model response to controlled changes in the sustainability-state variables. Notably, the moderate scenario increase of 6.8% is relatively close to the uncertainty boundary indicated by the predictive error and is therefore interpreted conservatively. The 95% bootstrap confidence intervals reported in Table 12 are non-overlapping across the reported scenario-level mean predicted yields, providing supporting evidence of separation among the scenario-level predictions; however, these intervals should not be interpreted as formal confidence intervals for the differences between scenarios. Nevertheless, these results represent model-based sensitivity estimates for comparative scenario exploration rather than causal forecasts or experimentally verified yield gains. The simulated trajectories should therefore be interpreted as model-based responses to the specified perturbations. The perturbation effects are evaluated on a paired district-year basis, such that each perturbed prediction is compared with its corresponding baseline prediction for the same observation. Consequently, the uncertainty assessment reflects the variability of the estimated scenario response across the evaluated observations rather than treating the baseline and perturbed predictions as independent samples. This paired assessment complements the bootstrap confidence intervals reported in Table 11 and provides a more direct basis for interpreting the magnitude and stability of the perturbation-induced responses.

The progressive increase in predicted yield demonstrates that the Digital Twin can be used to explore how changes in virtual sustainability states propagate through the trained forecasting model. These values represent model-based counterfactual sensitivity estimates and should not be interpreted as guaranteed real-world intervention effects. Their primary value is to support comparative "what-if" analysis and to identify potentially beneficial configurations for subsequent empirical or policy evaluation.

3.7. Retrospective Temporal Twin Validation

The temporal fidelity of TSAFI-DT was assessed through retrospective sequential replay, in which predictions generated using information available at each historical time

point were compared with subsequently observed yields. Figure 6 shows that the predicted trajectory generally follows the direction and temporal variation in the observed yield series.

The mean forecast discrepancy remained below 0.18 t/ha, consistent with the overall HB-XGBoost RMSE of 0.171 t/ha. The close temporal correspondence indicates that the virtual representation can reproduce the principal year-to-year variations in observed productivity under retrospective out-of-sample evaluation.

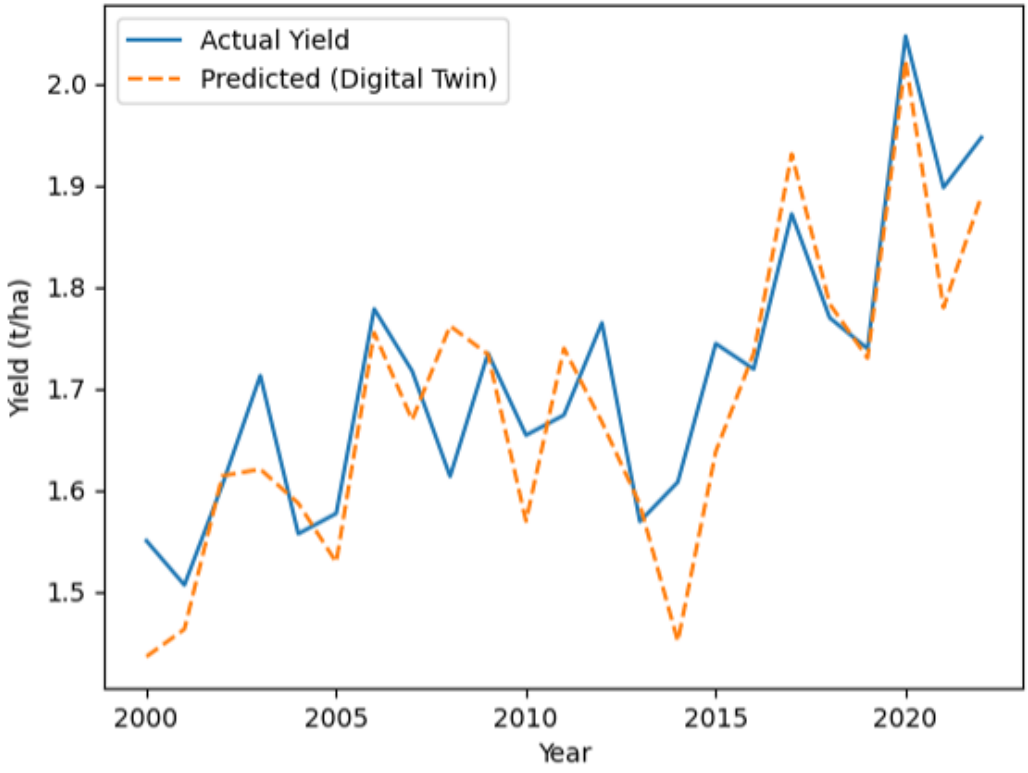

Figure 6. Temporal comparison of observed and TSAFI-DT-predicted groundnut yield under retrospective sequential validation.

These findings support the fidelity of TSAFI-DT as a retrospective data-driven Digital Twin prototype. They should not, however, be interpreted as validation of a continuously synchronized physical Digital Twin, since the current implementation operates on historical APY observations rather than live sensor or remote-sensing streams.

3.8. Predictive Ablation Analysis

Ablation analysis was performed only for components that directly contribute to the predictive model. In accordance with the methodological design, the counterfactual and fixed-effects analytical modules were not removed as predictive components because they operate downstream of yield forecasting and therefore do not directly determine predictive RMSE.

The inclusion of the lag-based benchmarks provides a more stringent temporal comparison and confirms that the predictive advantage of HB-XGBoost extends beyond simple persistence based on previous-year yield and cultivated area and is shown in Table 13. The lag-yield-based benchmarks provide a direct assessment of the predictive information already contained in recent productivity and cultivated-area history. Comparison with the proposed HB-XGBoost model indicates whether the inclusion of eRAI provides additional predictive value beyond these conventional temporal predictors. The results therefore

support the interpretation of eRAI as a complementary sustainability-state representation rather than as a replacement for historical yield information.

Table 13. Predictive performance comparison of lag-based benchmarks and evaluated models.

Model	MAE	RMSE	R^2
Lag-Yield	0.290	0.375	0.60
Lag-Yield + Area	0.255	0.325	0.70
Linear Regression	0.221	0.291	0.78
XGBoost	0.184	0.239	0.85
GA+XGBoost	0.149	0.204	0.89
HB-XGBoost	**0.128**	**0.171**	**0.92**

Table 14 compares the complete forecasting configuration with a variant in which the eRAI-based sustainability-state representation was removed. The complete model achieved an RMSE of 0.171 t/ha, whereas removal of eRAI increased RMSE to 0.195 t/ha, corresponding to a 14.3% deterioration.

Table 14. Predictive ablation analysis of the sustainability-state representation.

Model Variant	RMSE (t/ha)	Increase (%)
Full predictive TSAFI-DT	0.171	—
Without eRAI features	0.195	+14.3%

The increase in prediction error after removing eRAI indicates that the integrated sustainability-state representation contributes useful information beyond the remaining temporal, spatial, and historical yield predictors. This observation is consistent with the SHAP analysis, in which eRAI exhibited the largest overall feature contribution.

The counterfactual and fixed-effects components are evaluated separately through counterfactual fit, intervention-associated estimates, and scenario interpretability rather than through predictive RMSE. This separation avoids attributing forecasting improvements to downstream analytical modules that do not participate directly in model prediction. To further distinguish the contribution of eRAI from information contained in historical productivity, an additional lag-yield-based benchmark was considered. Specifically, a lag-yield-only [and/or lag-yield-plus-area] specification was evaluated using the same expanding-window temporal validation procedure. This comparison provides a stronger temporal benchmark against which the incremental predictive contribution of eRAI can be assessed.

Overall, the experimental findings demonstrate three complementary capabilities of TSAFI-DT. First, the framework achieves accurate temporal yield forecasting, with an RMSE of 0.171 t/ha and $R^2 = 0.92$. Second, the eRAI representation captures heterogeneous and persistent sustainability-state dynamics while improving predictive performance. Third, the Digital Twin layer enables retrospective state updating and model-based scenario exploration, with simulated yield changes of up to 12.4% under the evaluated sustainability-state perturbations. Together, these results support TSAFI-DT as a data-driven prototype for sustainability-aware agricultural forecasting and decision-support analysis.

3.9. Limitation of the Work

The Digital Twin scenario simulations also inherit the uncertainty of the underlying predictive model. Although the bootstrap analysis provides confidence intervals for the simulated scenario responses, these intervals do not eliminate uncertainty associated with

model specification, extrapolation beyond the observed sustainability-state space, or unobserved agricultural factors. In particular, the estimated response under the moderate scenario is relatively close to the RMSE uncertainty boundary and should therefore be interpreted cautiously. The simulated trajectories are intended for comparative sensitivity analysis rather than direct causal forecasting. Field-level empirical validation of the simulated trajectories is required before these scenario estimates can be considered for operational agricultural decision-making.

The empirical validation of TSAFI-DT is currently restricted to district-level groundnut yield data from the Indian APY setting. Therefore, the reported predictive performance and sustainability-state patterns should not be assumed to generalize directly to other crops, regions, or spatial resolutions. Transferability to other crops and agro-climatic regions would require recalibration of the sustainability indicators, eRAI components and weights, and model parameters using context-specific data. Extension to finer spatial scales, such as sub-district or farm level, would additionally require sufficiently granular, consistent, and reliable data. Multi-crop, multi-region, and finer-resolution validation therefore remains an important direction for future work.

4. Conclusions

Sustainable agricultural decision-making requires integrated frameworks capable of combining accurate yield forecasting with interpretable sustainability assessment. TSAFI-DT addresses this need through spatiotemporal reconstruction, structural-change detection, the Extended Regenerative Agriculture Index (eRAI), hierarchical Bayesian-optimized XGBoost, explainable AI, counterfactual assessment, fixed-effects analysis, and scenario simulation. HB-XGBoost achieved an RMSE of 0.171 t/ha and an R^2 of 0.92, outperforming the evaluated baseline models. Sustainability-state analysis revealed heterogeneous district-level transitions, while counterfactual and fixed-effects analyses provided complementary intervention-associated and adjusted associational evidence. Model-based scenario simulations estimated predicted yield increases of 6.8% and 12.4% under moderate and high sustainability-state perturbations, respectively; these estimates represent model sensitivity rather than guaranteed causal effects. Retrospective sequential validation further demonstrated temporal forecasting capability. Future work will integrate real-time IoT, weather, soil, and remote-sensing observations to enable continuously synchronized Digital Twin applications and prospective field-based validation. Future validation across multiple crops, regions, and finer spatial scales is required to establish the broader transferability of the TSAFI-DT framework.

Author Contributions: Conceptualization, R.R.N. and T.B.; methodology, R.R.N. and S.Y.; software, R.R.N.; validation, R.R.N., T.B., and S.Y.; formal analysis, R.R.N.; investigation, R.R.N.; data curation, R.R.N.; visualization, R.R.N.; writing—original draft preparation, R.R.N.; writing—review and editing, T.B., S.Y., and S.F.B.A.R.; supervision, T.B., S.Y., and S.F.B.A.R.; project administration, T.B. and S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are publicly available from the Directorate of Economics and Statistics, Department of Agriculture and Farmers Welfare, Government of India. The Agricultural Production Yield (APY) dataset can be directly accessed from: https://indiadataportal.com/p/area-production-yield-apy/r/moafw-area_production_yield-dt-sn-qzf, accessed on 6 September 2026. The processed data and code supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: All the authors acknowledge the use of the AI-based tool Grammarly for language editing assistance only. The tool did not contribute to the scientific content, analysis, or conclusions of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

APY	Area, Production, and Yield
CDI	Crop Diversity Index
DT	Digital Twin
eRAI	Extended Regenerative Agriculture Index
GA	Genetic Algorithm
HB-XGBoost	Hierarchical Bayesian-Optimized Extreme Gradient Boosting
IoT	Internet of Things
LUE	Land-Use Efficiency
MAE	Mean Absolute Error
PELT	Pruned Exact Linear Time
RMSE	Root Mean Squared Error
SC	Synthetic Control
SHAP	SHapley Additive exPlanations
SHP	Soil Health Proxy
TSAFI-DT	Temporal-Spatial Agricultural Forecasting and Intelligence Digital Twin
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting
YTM	Yield Trend Momentum

References

1. Botero-Valencia, J.; García-Pineda, V.; Valencia-Arias, A.; Valencia, J.; Reyes-Vera, E.; Mejia-Herrera, M.; Hernández-García, R. Machine learning in sustainable agriculture: Systematic review and research perspectives. *Agriculture* **2025**, *15*, 377. [CrossRef]
2. Araujo, S.O.; Peres, R.S.; Ramalho, J.C.; Lidon, F.; Barata, J. Machine learning applications in agriculture: Current trends, challenges, and future perspectives. *Agronomy* **2023**, *13*, 2976. [CrossRef]
3. Ullah, N.; Martínez-Álvarez, F.; De Falco, I.; Sannino, G. Optimizing Deep Learning for Cotton Leaf Disease Detection Using Meta-Heuristic Feature Selection Algorithms. In *Proceedings of the 2025 20th Conference on Computer Science and Intelligence Systems (FedCSIS)*; IEEE: Piscataway, NJ, USA, 2025; pp. 605–614.
4. Chawla, T.; Gahlawat, T.; Thakur, T. A Big Data Analytics–Based Architecture for Smart Farming. In *Optimizing AI Applications for Sustainable Agriculture*; Scrivener Publishing LLC.: New York, NY, USA, 2025; pp. 399–416.
5. Tao, F.; Zhang, M.; Liu, Y.; Nee, A. Digital Twins and Cyber–Physical Systems toward Smart Manufacturing and Industry 4.0: Correlation and Comparison. *Engineering* **2019**, *5*, 653–661. [CrossRef]
6. Fuller, A.; Fan, Z.; Day, C.; Barlow, C. Digital Twin: Enabling Technologies, Challenges and Open Research. *IEEE Access* **2020**, *8*, 108952–108971. [CrossRef]
7. Jones, D.; Snider, C.; Nassehi, A.; Yon, J.; Hicks, B. Characterising the Digital Twin: A systematic literature review. *Cirp J. Manuf. Sci. Technol.* **2020**, *29*, 36–52. [CrossRef]
8. Sheoran, S. Digital Twins for the Agriculture Sector. In *Blockchain and Digital Twin Applications in Smart Agriculture*; Auerbach Publications: Boca Raton, FL, USA, 2025; pp. 237–250.
9. Santhoshini, B.; Shanmugam, M. Smart Agriculture and Digital Twins: A Vision of Sustainability. In *Blockchain and Digital Twin Applications in Smart Agriculture*; Auerbach Publications: Boca Raton, FL, USA, 2025; pp. 17–28.
10. Pylianidis, C.; Osinga, S.; Athanasiadis, I.N. Introducing Digital Twins to Agriculture. *Comput. Electron. Agric.* **2021**, *184*, 105942. [CrossRef]
11. Verdouw, C.; Tekinerdogan, B.; Beulens, A.; Wolfert, S. Digital Twins in Smart Farming. *Agric. Syst.* **2021**, *189*, 103046. [CrossRef]
12. Elijah, O.; Rahman, T.; Orikumhi, I.; Leow, C.; Hindia, M. An Overview of Internet of Things (IoT) and Data Analytics in Agriculture. *IEEE Internet Things J.* **2018**, *5*, 3758–3773. [CrossRef]
13. Afzal, M.; Saeed, I.A.; Sohail, M.N.; Saad, M.H.M.; Sarker, M.R. IoT-enabled adaptive watering system with ARIMA-based soil moisture prediction for smart agriculture. *IEEE Access* **2025**, *13*, 27714–27728. [CrossRef]

14. Purcell, W.; Neubauer, T. Digital Twins in Agriculture: A State-of-the-Art Review. *Smart Agric. Technol.* **2023**, *3*, 100094. [CrossRef]
15. You, J.; Li, X.; Low, M.; Lobell, D.; Ermon, S. Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. *Proc. AAAI Conf. Artif. Intell.* **2017**, *31*, 4559–4565. [CrossRef]
16. Khaki, S.; Wang, L. Crop Yield Prediction Using Deep Neural Networks. *Front. Plant Sci.* **2019**, *10*, 621. [CrossRef] [PubMed]
17. Cesco, S.; Sambo, P.; Borin, M.; Basso, B.; Orzes, G.; Mazzetto, F. Smart Agriculture and Digital Twins: Applications and Challenges in a Vision of Sustainability. *Eur. J. Agron.* **2023**, *146*, 126809. [CrossRef]
18. Nair, R.R.; Babu, T.; Nayak, D.; Yogarayan, S.A.; Razak, S.F.A. Field Intelligence for Regenerative Agriculture in Predictive Sustainable Groundnut Farming in India. *Procedia Comput. Sci.* **2026**, *283*, 3730–3739. [CrossRef]
19. Wang, X.; Ma, Y.; Xu, Y.; Huang, Q.; Yang, Z.; Zhang, Z. Learning county from pixels: Corn yield prediction with attention-weighted multiple instance learning. *Int. J. Remote Sens.* **2025**, *46*, 2815–2845. [CrossRef]
20. Ayanlade, T.T.; Van der Laan, L.; Liu, Q.; Gangopadhyay, T.; Shook, J.; Singh, A.; Ganapathysubramanian, B.; Sarkar, S.; Singh, A.K. Transformer model to determine spatio-temporal relationships of variables, and interpretability for soybean seed yield, oil, and protein prediction. *Front. Artif. Intell.* **2026**, *9*, 1750108. [CrossRef] [PubMed]
21. Manocha, A.; Sood, S.K.; Bhatia, M. IoT-Digital Twin-Inspired Smart Irrigation Approach for Optimal Water Utilization. *Sustain. Comput. Inform. Syst.* **2024**, *41*, 100947. [CrossRef]
22. Abadie, A.; Diamond, A.; Hainmueller, J. Synthetic Control Methods for Comparative Case Studies. *J. Am. Stat. Assoc.* **2010**, *105*, 493–505. [CrossRef]
23. Angrist, J.D.; Pischke, J.S. *Mostly Harmless Econometrics: An Empiricist's Companion*; Princeton University Press: Princeton, NJ, USA, 2008.
24. Callaway, B.; Sant'Anna, P.H. Difference-in-Differences with Multiple Time Periods. *J. Econom.* **2021**, *225*, 200–230. [CrossRef]
25. Srichaiyan, P.; Tippayawong, K.Y.; Boonprasope, A. Forecasting soybean futures prices with adaptive AI Models. *IEEE Access* **2025**, *13*, 48239–48256. [CrossRef]
26. Rayalu, M. Optimizing rice yield prediction: A policy-oriented approach to agricultural sustainability. *Results Eng.* **2025**, *28*, 107844. [CrossRef]
27. Escribà-Gelonch, M.; Liang, S.; van Schalkwyk, P.; Fisk, I.; Long, N.V.D.; Hessel, V. Digital twins in agriculture: Orchestration and applications. *J. Agric. Food Chem.* **2024**, *72*, 10737–10752. [CrossRef] [PubMed]
28. Kim, S.; Heo, S. An agricultural digital twin for mandarins demonstrates the potential for individualized agriculture. *Nat. Commun.* **2024**, *15*, 1561. [CrossRef] [PubMed]
29. Karaarslan, E.; Aydin, Ö.; Cali, Ü.; Challenger, M. *Digital Twin Driven Intelligent Systems and Emerging Metaverse*; Springer: Berlin/Heidelberg, Germany, 2023.
30. Ai, L.; Ziehl, P. Advances in digital twin technology in industry: A review of applications, challenges, and standardization. *J. Intell. Constr.* **2025**, *3*, 9180083. [CrossRef]
31. Zohdi, T.I. A Machine-Learning Enabled Digital-Twin Framework for Next Generation Precision Agriculture and Forestry. *Comput. Methods Appl. Mech. Eng.* **2024**, *431*, 117250. [CrossRef]
32. Nie, J.; Wang, Y.; Li, Y.; Chao, X. Artificial intelligence and digital twins in sustainable agriculture and forestry: A survey. *Turk. J. Agric. For.* **2022**, *46*, 642–661. [CrossRef]
33. Karanam, S.A.K.; Hartman, N.W. A systematic review of digital twin (DT) and virtual learning environments (VLE) for smart manufacturing education. *Manuf. Lett.* **2025**, *44*, 1597–1608. [CrossRef]
34. Smith, J.C.; Ghosh, A.; Hijmans, R.J. Agricultural intensification was associated with crop diversification in India (1947–2014). *PLoS ONE* **2019**, *14*, e0225555. [CrossRef] [PubMed]
35. Griffiths, B.S.; Faber, J.; Bloem, J. Applying Soil Health Indicators to Encourage Sustainable Soil Use: The Transition from Scientific Study to Practical Application. *Sustainability* **2018**, *10*, 3021. [CrossRef]
36. Lin, H.C.; Hülsbergen, K.J. A new method for analyzing agricultural land-use efficiency, and its application in organic and conventional farming systems in southern Germany. *Eur. J. Agron.* **2017**, *83*, 15–27. [CrossRef]
37. Pankhurst, C.E.; Doube, B.M.; Gupta, V.V.S.R. Biological indicators of soil health: Synthesis. In *Biological Indicators of Soil Health*; CAB International: Wallingford, UK, 1997; pp. 419–435.
38. Nayak, D.; Nair, R.R.; Babu, T.; Rajamani, K. XGBoost-Based Framework for Robust Crop Yield Prediction under Uncertainty Using Fuzzy Logic and Genetic Optimization. *Procedia Comput. Sci.* **2026**, *282*, 539–548. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.