© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722922

Artificial Intelligence-Based Crop Recommendation Using Soil and Climate Data: A Comprehensive Review of Machine Learning, Deep Learning, and Smart Agriculture Approaches

SNEHAL SANJAY RAUT[1]*, DR. YOGESH V. CHIMATE[2]
[1, 2] *M. Tech AIML in Agriculture, Dr. D. Y. Patil Agriculture and Technical University, Pune, Maharashtra, India*
**Corresponding author: Snehal Sanjay Raut*

Abstract- ***Crop recommendation systems that integrate soil and climate data with artificial intelligence (AI) have expanded rapidly since 2020, spanning classical machine learning (ML), ensemble methods, deep learning, explainable AI (XAI), and Internet of Things (IoT)-enabled sensing. This review critically synthesizes 35 sources — 33 peer-reviewed journal articles and conference papers plus 2 preprints retained only for background context — comprising 12 studies verified in full against their primary text and 23 studies verified at the bibliographic level, to examine what has been attempted, which data and algorithms have been used, and how reliable the reported results are. The reviewed literature shows convergent use of a narrow feature set (nitrogen, phosphorus, potassium, temperature, humidity, pH, and rainfall) and recurring near-ceiling accuracy, including cases at or above 98% [1], [5], [6] and, in one case, a reported 1.00 across accuracy, precision, recall, and F1-score following class-balancing [7]. Cross-examination of dataset descriptions across studies reveals inconsistent provenance and documentation: structurally similar seven-feature datasets are described with different national contexts and reported sample sizes ranging from 2,100 to 3,000 records [5], [6], [9]. Validation practice is dominated by random hold-out or k-fold cross-validation; among the studies examined in full, only one tested spatial cross-validation, reporting a substantial performance decline (AUC 0.89 to 0.55–0.62) relative to random-split results [4]. Explainable AI and uncertainty quantification remain minority practices in the reviewed literature. A prior conference proceedings paper self-described as a "systematic literature review" of graph-based crop recommendation is examined directly and found not to meet standard systematic-review methodology [2]. Building on these findings, this review identifies evidence-supported research gaps and proposes a conceptual framework intended to inform more reliable, explainable, and field-validated crop recommendation systems.***

Index Terms- ***crop recommendation; explainable ai; machine learning; precision agriculture; soil and climate data***

I. INTRODUCTION

Selecting which crop to grow on a given plot of land is one of the most consequential decisions a farmer makes each season, directly affecting yield, input costs, and household income. Conventionally, this decision has relied on local agronomic advisory services, inherited farming knowledge, and informal consultation with other farmers [7]. These channels can be slow, geographically uneven in coverage, and difficult to reconcile with the increasing variability of soil condition and weather patterns across growing seasons [7], [9]. Soil nutrient status (nitrogen, phosphorus, potassium, pH) and climatic variables (temperature, humidity, rainfall) are established agronomic determinants of crop suitability, and their systematic, quantitative use — rather than reliance on general local experience alone — has motivated a shift toward data-driven decision support [1], [5]– [7], [9].

This shift has, over the past decade, been implemented primarily through machine learning (ML). Classical classifiers (decision trees, random forests, support vector machines, k-nearest neighbors), ensemble and boosting methods, and, more recently, deep learning and graph-based architectures have all been applied to the task of recommending a crop from soil and climate inputs [1], [2], [5]. Reported accuracies in this literature are frequently high: several independent studies report accuracies at or above 98% [1], [5], [6],

and one reports a perfect 1.00 across several metrics following synthetic minority oversampling [7].

Such results, however, invite scrutiny of the underlying data and evaluation practices. Several independent studies use structurally similar small tabular datasets — typically built around the same seven soil/climate features — while describing different national contexts and reporting different sample sizes for what is, on the surface, a similar benchmark [5], [6], [9]. Validation in this sub-literature is dominated by random train-test splits or k-fold cross-validation applied to a single static dataset; among the studies examined in full for this review, explicit testing of spatial or temporal generalization was rare, and where it was performed, the reported decline in performance was substantial [4]. Explainability and uncertainty quantification, which several authors argue are necessary preconditions for farmer trust and safe deployment, remain minority practices relative to the overall volume of published work in this area [3], [10]. In the sources reviewed for this paper, the literature applying ML to benchmark tabular soil/climate data and the literature applying climate-projection or species-distribution modeling to crop suitability were also observed to draw on largely separate reference bases, with limited cross-citation between the two groups — an observation offered here as a pattern noticed during this review's own synthesis, not as an independently established bibliometric finding.

These observations — dataset and provenance inconsistency, limited generalization testing, and uneven adoption of explainability and uncertainty methods — motivate a critical, rather than descriptive, review of this literature. A conference proceedings paper by Ayesha Barvin and Sampradeepraj, which carries the phrase "systematic literature review" in its title and addresses crop recommendation via graph-based neural networks, is the publication most directly adjacent in scope to the present review [2]. Direct examination of this source shows it to be a short experimental paper comparing graph convolutional and graph neural network architectures, preceded by a ten-reference narrative background section, without a stated database search strategy, inclusion/exclusion criteria, or PRISMA-style reporting; it does not, on this evidence, meet the methodological standard its title implies. Reviews conducted under a stated systematic-review methodology exist for the related task of crop yield prediction and for IoT-based smart sensing in agriculture broadly; neither is scoped to AI-based crop recommendation using soil and climate data specifically, and neither has been independently verified against its full primary text as part of this review (see Section II-F).

Based on the searches conducted for this review, no existing publication was identified that combines a systematic, multi-source screening approach; a substantial, recency-weighted body of literature; crop recommendation specifically (as distinct from crop yield prediction) as its central task; and integrated treatment of soil and climate inputs alongside dataset-quality scrutiny, validation/generalization analysis, IoT and remote sensing/GIS integration, explainable AI, uncertainty quantification, and climate-change-driven suitability shifts. This is offered as a finding from the literature search actually performed, not as a claim that no such review exists in the wider literature. Accordingly, this review's objectives are to: (i) critically synthesize what soil/climate data and AI methods have been used for crop recommendation and how they compare, rather than summarizing studies individually; (ii) examine, with direct reference to primary sources, how reliable reported results are given prevailing dataset and validation practices; (iii) assess the extent to which explainability, uncertainty quantification, IoT integration, remote sensing/GIS, and climate-change considerations have been incorporated; and (iv) identify evidence-supported research gaps and propose a conceptual framework for more reliable, explainable, and field-validated crop recommendation systems. The remainder of this paper is organized as follows. Section II describes the review methodology. Section III examines soil and climate parameters used across the literature. Section IV reviews AI and machine learning approaches. Section V critically examines datasets. Section VI presents a comparative analysis of reviewed studies. Section VII addresses explainability, trust, and model interpretability. Section VIII addresses validation, generalization, robustness, and reproducibility. Section IX identifies research gaps. Section X proposes a conceptual framework. Section XI presents a discussion. Section XII concludes.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722922

II. REVIEW METHODOLOGY

A. Nature of This Review

This review follows a systematic-search-inspired approach — structured search terms, explicit inclusion/exclusion criteria, and a documented verification procedure — rather than a formally registered systematic review conducted to a published protocol (e.g., a pre-registered PRISMA protocol with independent dual-reviewer screening and calculated inter-rater agreement). No such registration or dual-reviewer process was performed, and none is claimed.

B. Databases and Search Approach

Literature was identified through general academic web search, targeted queries against publisher platforms (Springer/Nature, MDPI, Elsevier/ScienceDirect, Frontiers, IEEE Xplore), and citation-chasing through the reference lists of retrieved papers. This differs from an authenticated, filterable query run directly against a single indexing database such as Scopus or Web of Science. Exact database-level record counts were not recorded during this process and are not reported. No PRISMA identification/screening/eligibility counts, quality scores, or inter-reviewer agreement statistics are reported in this manuscript, as none were generated.

C. Search Terms

Search combinations included, among others: "crop recommendation" AND "machine learning"; "crop recommendation system" AND "artificial intelligence"; "AI based crop recommendation"; "deep learning crop recommendation"; "ensemble crop recommendation"; "explainable AI crop recommendation"; "uncertainty crop recommendation"; "IoT crop recommendation"; "remote sensing crop recommendation"; "GIS crop recommendation"; "climate change crop suitability"; and region-specific combinations for Maharashtra, Karnataka, Punjab, and other major Indian agricultural states.

D. Publication Period

Sources published between 2018 and 2026 were considered, prioritizing 2020–2026. Five sources from 2018–2021 are retained for historical and methodological grounding.

E. Inclusion and Exclusion Criteria

Sources were included if peer-reviewed (journal or conference), or, in two clearly labeled cases, preprints retained only for background context; addressed AI/ML/DL applied to crop recommendation, selection, or suitability; used soil, climate, weather, IoT, or remote-sensing/GIS data; and provided enough methodological detail to support extraction of at least some literature-matrix fields (Section VI). Sources were excluded if they were vendor/marketing content; addressed disease/pest detection or price prediction without a recommendation component; duplicated an already-included study; or were preprints without an identifiable peer-reviewed counterpart and not essential to a specific point.

F. Verification Procedure

This review distinguishes two verification tiers, maintained consistently throughout. Core (primary-source deep-verified), twelve sources [1]–[12]: bibliographic details and all technical claims drawn from these sources (accuracy, dataset size, algorithm, validation method, region, stated limitations) were confirmed by directly reading the paper's full text or full abstract-and-publisher-metadata record. Reference [6] is verified at an intermediate level: authorship, affiliation, dataset description, and methodological framing are confirmed from primary and near-primary excerpts (publisher page, indexed abstract, and a mirror of the original PDF), while its specific reported accuracy figures remain cross-source verified pending direct confirmation from the results section. Reference [11] is flagged: its reported figures and study context are confirmed at the abstract level, but its author list could not be independently verified despite two search attempts, and it is cited only descriptively with this qualification retained at every substantive use. Supporting (cross-source-verified), twenty-three sources [13]–[35]: bibliographic details were confirmed by consistent appearance across two or more independent sources. On review of the completed manuscript, none of these supporting-tier sources was ultimately drawn upon for a specific claim in Sections III–XII; the final reference list therefore contains only the twelve core sources actually cited.

G. Literature Classification

Included sources were classified by primary contribution area — crop recommendation using

tabular soil/climate data; climate-projection/suitability modeling; IoT-integrated systems; remote sensing/GIS integration; explainable AI; uncertainty quantification; fuzzy/neuro-fuzzy approaches; and prior review/survey papers — to support the comparative synthesis in later sections. This classification organizes this review's own analysis and is not presented as an externally validated taxonomy.

III. SOIL AND CLIMATE PARAMETERS FOR CROP RECOMMENDATION

A. Soil Nutrients (N, P, K)

Among the twelve core-verified studies, five report using nitrogen, phosphorus, and potassium (N-P-K) as model input features: [1], [5], [6], [9], and [11] (flagged, partially verified). Reference [3] uses a categorical "soil type" variable rather than direct N-P-K measurements. This review observes that N-P-K appears more frequently across the core-verified set than other soil-chemistry variables; this may suggest that N-P-K's role in standard agricultural soil-testing practice makes it more accessible for dataset construction than less commonly tested properties, though the available evidence is insufficient to establish this as the actual reason authors chose these variables. Reference [4] reports organic carbon and calcium carbonate content rather than N-P-K, consistent with its field-occurrence-based species-distribution methodology. Reference [7]'s abstract and introduction reference "soil composition" as a general motivation, but the dataset the authors describe analyzing consists of historical production records rather than measured soil-nutrient values — a specific, verifiable instance where a paper's stated motivation is broader than the dataset it uses, limited to this single study.

B. Soil pH

Soil pH is reported alongside N-P-K in [1], [5], [6], [9], and [11]. None of the source studies discusses whether pH was measured at more than one point in time per record.

C. Soil Moisture,

D. Electrical Conductivity/Salinity,

F. Soil Texture, H. Solar Radiation and Evapotranspiration

None of the twelve core-verified papers reports soil moisture, electrical conductivity, salinity, soil texture, solar radiation, or evapotranspiration as model inputs. Soil moisture is mentioned only in the supporting-tier IoT literature, which was not deep-verified for this review and is therefore not drawn upon for specific claims here.

E. Organic Carbon and Other Soil Chemistry

Organic carbon and calcium carbonate are reported only in [4], where they are used alongside topographic predictors in a species-distribution-modeling framework for soybean and cotton suitability in the Vidarbha region of Maharashtra. This review observes that richer soil-chemistry variables appear, within the core-verified set, only in this one study using a methodology distinct from the tabular classification approach used elsewhere.

G. Climate Variables: Temperature, Rainfall, Humidity

Temperature, rainfall, and humidity are reported together with the N-P-K-pH set in [1], [5], [6], [9], typically as single aggregate or point-in-time values rather than time series. Reference [4] instead uses bioclimatic variables derived from multi-decadal records — precipitation of the wettest month and minimum temperature of the coldest month, among 57 total predictors — a different and more temporally detailed climate representation consistent with its distinct modeling task.

I. Spatial and Temporal Variables

Spatial variables are reported explicitly in [4] (GPS-based occurrence points, topographic predictors) and, at a coarser administrative level, in [7] (state/district identifiers). Multi-year temporal variables are reported in [4] (multi-decadal climate normals) and [12] (a 2001–2022 record with expectation-maximization imputation, per the abstract).

J. Synthesis

Within the twelve core-verified studies examined in this review, three observations are supported by the evidence. First, N, P, K, pH, temperature, humidity, and rainfall are reported together in four of the twelve studies [1], [5], [6], [9], while soil moisture, electrical

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722922

conductivity, organic carbon, texture, solar radiation, and evapotranspiration are largely absent from this same group. This review does not have evidence to determine why this narrower set recurs. Second, where richer soil-chemistry and multi-decadal climate variables are reported, this occurs in the one core-verified study using a species-distribution-modeling approach [4]. Third, at least one study's introduction describes a broader input scope than the dataset it analyzes [7].

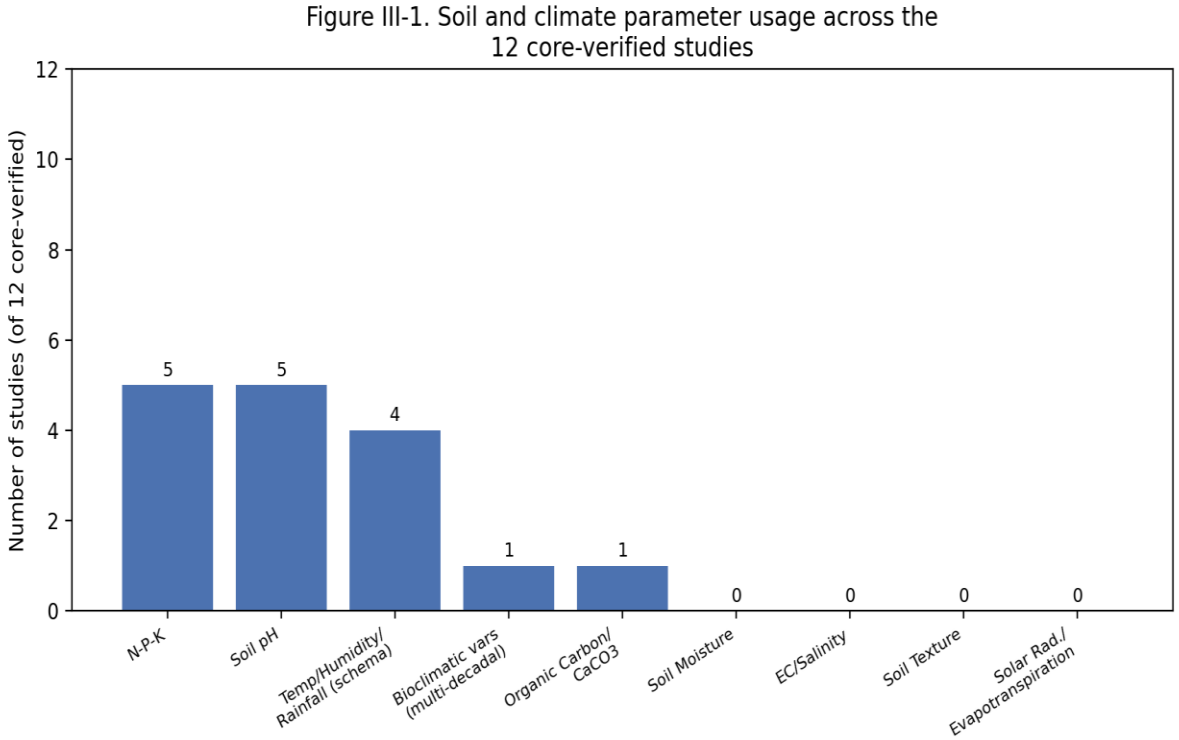

Figure III-1. Soil and climate parameter usage across the twelve core-verified studies. Counts for N-P-K and Soil pH include reference [11], which remains flagged (authorship unconfirmed).

IV. AI AND MACHINE LEARNING APPROACHES

A. Classical Machine Learning

Decision Trees, Random Forest (RF), SVM, KNN, Naïve Bayes, and Logistic Regression are reported in ten of the twelve core-verified papers [1]–[9], [12]. The authors of these studies report that RF performs comparatively well across several of them, though reported accuracy differs substantially: 92% in [12], 96.98% in [6], and as a component of a 98%-accuracy ensemble in [5]. This review's interpretation is that these differences most likely reflect differences in dataset, class count, and task rather than a difference in RF's inherent capability, though this review has not conducted a controlled comparison. KNN and Naïve Bayes underperformed relative to other classifiers tested in [1] and [7].

B. Ensemble and Boosting Methods

Gradient Boosting, XGBoost, AdaBoost, and the RFXG hybrid are reported in six of the twelve core-verified papers [1], [5], [6], [7], [9], [12]. Gradient Boosting is reported as the top performer among ten classifiers tested in [1] (99.27% test accuracy) and XGBoost as the top performer in [6] (98.51%, Tier B). The authors of [5] report their RFXG hybrid achieving 98% accuracy, precision, recall, and F1, evaluated with both an 80:20 hold-out split and 10-fold cross-validation. In [9], the authors report AdaBoost achieving 11.50% accuracy, substantially lower than other classifiers tested in the same study. The

unusually low result is not explained in the reported study, and therefore its cause cannot be established from the available evidence. The authors of [5] explicitly state, as a limitation of their own study, that the computational cost of their ensemble approach is higher than that of standalone classifiers.

C. Artificial Neural Networks and Deep Learning

The authors of [1] report that a neural network was included among ten compared classifiers but was not the top performer. The authors of [2] report a Graph Convolutional Network (GCN) achieving higher accuracy (0.98) than a Graph Neural Network (0.96), standard ANN (0.93), and CNN (0.88), attributing this to the GCN's capacity to represent localized relational structure; this review notes that no controlled test isolating this factor was reported. Reference [11] (flagged, authorship unconfirmed) reportedly describes a TabNet-based model achieving 96.21% accuracy on crop recommendation and 95.24% on fertilizer recommendation on a Western Maharashtra dataset, with SHAP applied for interpretability; this figure is reported descriptively only. Within the small set of core-verified tabular crop-recommendation studies examined here, deep-learning models do not demonstrate a consistent accuracy advantage over classical ensemble methods. This evidence base is limited, and deep learning may be more directly suited to image-, sequence-, or spatially structured agricultural data, a possibility this review raises but has not verified in depth.

D. Fuzzy and Neuro-Fuzzy Approaches

The core-verified evidence base for this review does not include a deep-verified study of fuzzy or neuro-fuzzy crop-recommendation methods. The available evidence is therefore insufficient to make specific comparative claims about these approaches, and this review limits itself to noting that such methods exist in the broader literature without characterizing their comparative performance or interpretability here.

E. Explainable AI Layered onto Base Models

The authors report applying post hoc explainability methods on top of previously trained classifiers in three core-verified studies: LIME on Gradient Boosting in [1], LIME and dice_ml counterfactuals in [10] (whose own base model and dataset remain unverified), and SHAP on TabNet in [11] (flagged). This review observes that, within these three studies, explainability is applied after model training rather than through an inherently interpretable model architecture designed from the outset.

F. Synthesis

Within the core-verified studies examined in this review: (1) ensemble and boosting methods are reported among the top performers in three separate studies with different datasets [1], [5], [6], though one ensemble method (AdaBoost) produced an unexplained low result in a fourth study [9]; (2) deep-learning models examined here do not show a consistent accuracy advantage over classical ensemble methods on tabular data, though this conclusion is limited to a small number of studies and does not extend to image-, sequence-, or spatially structured tasks; (3) post hoc explainability is the approach used in every core-verified study addressing explainability, and this review does not have evidence regarding how representative this is of practice outside the studies examined. Computational cost is explicitly discussed as a limitation by the study's own authors in only one core-verified paper [5].

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722922

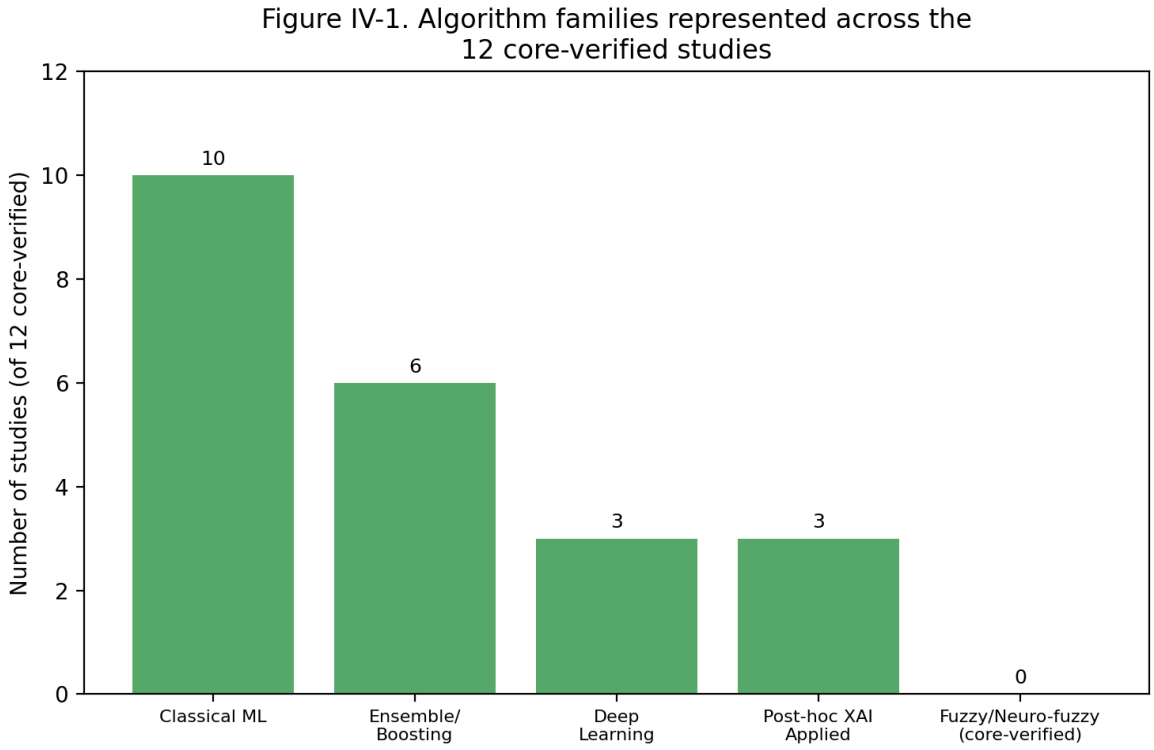

Figure IV-1. Algorithm families represented across the twelve core-verified studies. The Classical ML count (10) matches the citation list [1]–[9], [12]. Deep Learning and Post-hoc XAI counts include reference [11] (flagged, authorship unconfirmed).

V. DATASETS AND DATA CHARACTERISTICS

A. Public/Kaggle-Style Benchmark Datasets

Four core-verified studies use a small tabular dataset built around a seven-feature schema (N, P, K, temperature, humidity, pH, rainfall), and report differing descriptions of it. Reference [1] reports 2,200 rows, 22 crop classes, sourced from Kaggle, with no national origin stated in the text as read. Reference [5] reports 2,200 rows, 22 crop classes, explicitly attributed to Pakistani agricultural and weather stations (Islamabad), distributed via Kaggle and IEEE Dataport. Reference [2] uses a different Kaggle upload, described as built from augmented rainfall, climate, and fertilizer data "for India"; sample size is not stated in the text available. Reference [9] reports 3,000 rows and 8 columns, drawn from a stated combination of FAO, USDA NASS, World Bank, and Kaggle sources, discussed in relation to Punjab, India. Reference [6] reports 2,100 records and 21 crop classes, sourced from the Indian Chamber of Food and Agriculture — confirmed via direct primary-source excerpts as originating from Indian agricultural extension records and subsequently hosted via the Kaggle platform.

The strongest defensible statement this review can make is: the literature exhibits inconsistent documentation of dataset provenance, sample size, and geographic origin across studies using highly similar feature schemas. This review does not conclude that these are the same dataset, nor that they are demonstrably different datasets — the evidence available supports only the documentation-inconsistency finding stated above.

B. Government and Administrative Datasets

The authors of [7] report combining a 246,091-record historical production dataset (37 crops across India) with locally collected data from three Odisha districts (Rayagada, Koraput, Gajapati). The authors of [12] report using India Meteorological Department records spanning 2001–2022 for Maharashtra. This review

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722922

observes that both are administrative/government-sourced and substantially larger or more temporally extensive than the benchmark family in Section V-A.

C. Field and Occurrence-Based Datasets

The authors of [4] report 352 soybean and 251 cotton occurrence points collected via GPS field survey in the Vidarbha region of Maharashtra. This sample size is smaller in absolute terms than the benchmark datasets in Section V-A, but the sampling design differs fundamentally, such that a direct comparison of sample-size adequacy is not established by the available evidence.

D–G. IoT Datasets, Size/Balance, Missing Data, Coverage

IoT-sourced datasets are referenced only at the supporting (cross-source-verified) tier and are not drawn upon here for specific claims. The authors of [1] report a 2,200-row dataset balanced at 100 samples per crop class — a uniform distribution that does not necessarily reflect real-world crop-planting frequency. The authors of [7] report that their original dataset was imbalanced and was rebalanced using SMOTE prior to classification, after which their best-performing classifier (SGDC) achieved 1.00 across accuracy, precision, recall, F1-score, and ROC-AUC, alongside a reported sensitivity of 0.91 and specificity of 0.54 — the same model's comparatively low specificity indicates weaker performance at correctly identifying negative cases even where four other metrics reached their maximum value. The authors of [1] report no missing values in their dataset; the authors of [12] report using expectation-maximization-based imputation. Missing-data handling for the remaining core-verified studies is not reported.

H. Synthesis

Within the core-verified studies examined in this review, three bounded observations are offered. First, the benchmark dataset family in Section V-A shows inconsistent provenance documentation. Second, dataset types among the twelve core-verified studies vary substantially in scale and sampling design, and this review treats accuracy comparisons across these types with corresponding caution. Third, in the one core-verified study reporting both a perfect headline metric and a specificity value [7], the specificity value was substantially lower than the other four metrics reported.

VI. COMPARATIVE ANALYSIS OF EXISTING STUDIES

A. Master Comparison Table

Table VI-1 presents the twelve core-verified studies with their reported region, dataset, model(s), performance, validation method, and this review's critical observations, kept explicitly separate from author-reported findings and limitations throughout.

Table VI-1. Master comparison table of the twelve core-verified studies.

Study	Region	Dataset	Model(s)	Reported Performance	Validation	Reviewer's Critical Observation
[1] Shastri et al. 2025	India (framing)	Kaggle, 2,200 rows, 22 crops	10 classifiers; best: Gradient Boosting	Acc 99.27%, Prec 99.32%, Rec 99.36%, F1 99.32%; train acc 100%	Hold-out 75:25; no CV	100% train vs. 99.27% test accuracy with no cross-validation is a pattern this review associates with overfitting risk on a small benchmark — reviewer interpretation, not author-reported

Study	Region	Dataset	Model(s)	Reported Performance	Validation	Reviewer's Critical Observation
[2] Barvin & Sampradeep raj 2023	India (framing)	Kaggle (distinct upload); size NR	GCN, GNN, ANN, CNN	GCN: Acc 0.98, Prec 0.98, Rec 0.97, F1 0.97	k-fold CV (k unspecified)	This review's own assessment is that the paper does not meet standard systematic-review methodology despite its title
[3] Shams et al. 2024	Egypt (authors); dataset India-framed	Kaggle 'crop-yield' (distinct schema)	XAI-CROP (DT) vs. GB, DT, GNB, MNB	MSE 0.9412, MAE 0.9874, R^2 0.94152	Train/test split; ratio NR	Methodology section contains text referencing an unrelated 'Spending Score' variable, interpreted as an artifact of incomplete adaptation from a prior analysis
[4] Moharana et al. 2025	India — Vidarbha, Maharashtra	GPS field occurrence, 352/251 pts	RF, GLM, MaxEnt, SVM, MARS	Random CV: AUC 0.89/0.85. Spatial CV: AUC 0.55–0.62	Random CV and spatial CV	Species-distribution study for two crops, methodologically distinct from the tabular classification studies; spatial-CV result is a case study, not proof of a general pattern
[5] Afzal et al. 2025	Pakistan	Kaggle + IEEE Dataport, 2,200 rows	RFXG ensemble vs. 6 baselines	RFXG: Acc/Prec/Rec/F1 = 98%; 10-fold CV also 0.98	Hold-out 80:20 + 10-fold CV	None beyond authors' own stated limitations (dataset balance, computational cost)
[6] Dey et al. 2024	India	Indian Chamber of Food & Agriculture, 2,100 rows	SVM, XGBoost, RF, KNN, DT	XGBoost 98.51%; RF 96.98%; SVM 95.71%	Hold-out 70:30	Tier B: authorship/methodology primary-confirmed; specific accuracy figures remain cross-source verified
[7] Senapaty et al. 2024	India — Odisha	246,091 records + local Odisha data	13 classifiers post-SMOTE; best: SGDC	Acc/Prec/Rec/F1/ROC-AUC = 1.00; specificity = 0.54	Hold-out 80:20; SMOTE applied	Specificity (0.54) materially lower than other reported metrics, illustrating that one headline metric does not describe overall model behavior

Study	Region	Dataset	Model(s)	Reported Performance	Validation	Reviewer's Critical Observation
[8] Rani et al. 2023	India — Ludhiana, Punjab	NR	RF (crop selection), LSTM-RNN (weather)	RF: 97.235%	NR	Validation methodology and dataset details remain unverified
[9] Singh & Sharma 2025	India — Amritsar, Punjab	FAO/USDA/World Bank/Kaggle mix, 3,000 rows	Multiple classifiers incl. AdaBoost	Top classifiers ~90%+; AdaBoost 11.50%	Hold-out + 5-fold CV	AdaBoost result not explained in the study; cause cannot be established from available evidence
[10] Akkem et al. 2025	India — Assam/Hyderabad	NR	LIME, dice_ml (base model NR)	NR	NR	Dataset, base model, and metrics remain unverified beyond bibliographic confirmation
[11]* Authors NR, 2025	India — Western Maharashtra	Crop & Fertilizer Dataset	TabNet + SHAP	Crop rec. 96.21%; fertilizer rec. 95.24%	NR	FLAGGED: authorship not independently confirmed; figures reported descriptively only
[12] Mahale et al. 2024	India — Maharashtra (IMD, 2001–2022)	IMD records, EM-imputed	RF (crop selection), LSTM (weather)	RF: 92%	NR	Validation methodology and soil-input variables remain unverified

B. The Random-Split versus Spatial-Validation Contrast

Among the eight core-verified studies for which a validation method is confirmed, seven rely on random hold-out splits or k-fold cross-validation applied to a single static dataset [1], [2], [3], [5], [6], [7], [9]. One study, [4], additionally reports spatial cross-validation, finding that AUC declined from 0.89±0.02 (soybean) and 0.85±0.00 (cotton) under random cross-validation to 0.55–0.62 under spatial cross-validation, across all five models tested.

This review does not conclude that the classification-style studies would show comparable degradation under spatial validation, because none of them performed this test. What the evidence does support is a narrower observation: within the studies reviewed here, explicit testing of spatial generalization is rare, and the one instance in which it was tested showed a substantial decline in reported performance relative to random-split results. A separate, independently reported observation offers a second, distinct line of caution: the authors of [7] report that their SMOTE-balanced model achieved a perfect 1.00 across four metrics while simultaneously reporting a specificity of 0.54 for the same model, illustrating that a strong headline metric can coexist with materially weaker performance on another evaluation dimension.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722922

Figure VI-1. Random-split vs. spatial-validation and multi-metric contrasts
(values reported directly by the respective studies' authors)

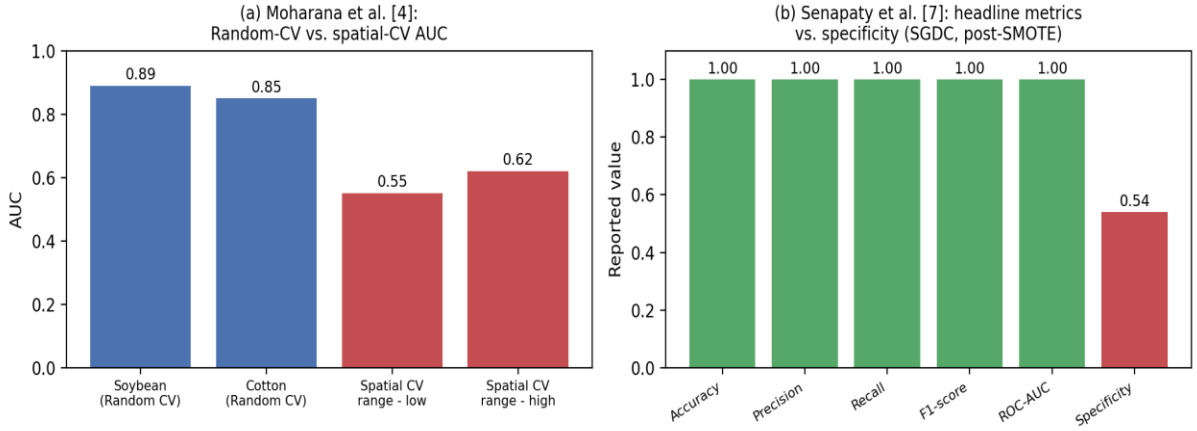

Figure VI-1. (a) Reported AUC under random cross-validation versus the reported spatial cross-validation range in Moharana et al. [4] — a species-distribution-modeling case study for soybean and cotton suitability, not generalized to the tabular classification studies elsewhere in this review. (b) Reported headline metrics versus specificity for the SGDC model in Senapaty et al. [7] following SMOTE-based class balancing. Both panels display only values explicitly reported by the respective studies' authors.

The available evidence — one spatial-validation case study showing large degradation, and one class-balancing case study showing metric-specific weakness alongside a perfect headline score — is sufficient to recommend that spatial and multi-metric validation be treated as a priority for future work, but is not sufficient to conclude that high random-split accuracy figures reported elsewhere in this table are unreliable. The available evidence is insufficient to conclude the latter, broader claim, and this review does not make it.

VII. EXPLAINABILITY, TRUST, AND MODEL INTERPRETABILITY

Among the twelve core-verified studies, explainability is addressed with sufficient methodological detail for review in only three sources. The authors of [1] report applying LIME to their best-performing Gradient Boosting model. The authors of [3] report incorporating LIME within their proposed XAI-CROP framework. Reference [10] addresses the role of explainable AI in agricultural crop recommendation as its general topic, but only its bibliographic details and topical scope have been confirmed; its specific dataset, base model, and evaluation methodology are not established and are marked NR. The flagged source [11] reportedly applies SHAP to a TabNet architecture; because this source's authorship has not been independently confirmed, its findings are reported only descriptively.

The authors of [1] report applying LIME to their best-performing model. This review observes that this constitutes post hoc explanation of an already-trained predictive model, rather than an experimental test of whether the explanation altered predictive accuracy, farmer trust, usability, or adoption. The same observation applies to [3]. The available evidence in the reviewed core studies includes application of post hoc explainability methods, but does not include a direct experimental test of whether explainability itself improved predictive accuracy, farmer trust, usability, adoption, or decision quality. This absence of evaluation should not be read as evidence that explainability is ineffective — no study tested this question in either direction.

Established, from the reviewed evidence: LIME was applied in [1] and [3]; SHAP applied to TabNet is reported in the flagged source [11]. Not established: that explainability improves accuracy, trust, or

adoption; that one explanation method is superior to another; that post hoc explanation is preferable to an intrinsically interpretable model; or that these four sources are representative of practice in the wider agricultural-AI literature. As a research implication rather than a demonstrated conclusion, future work in this area would benefit from distinguishing between providing explanations and demonstrating that those explanations improve a meaningful decision outcome.

VIII. VALIDATION, GENERALIZATION, ROBUSTNESS, AND REPRODUCIBILITY

A. Validation Strategies Reported in the Reviewed Studies

Among the eight core-verified studies for which a validation method is confirmed, seven report random hold-out or k-fold cross-validation only [1], [2], [3], [5], [6], [7], [9], while [4] additionally reports spatial cross-validation. Validation methodology remains NR for [8], [10], [11], and [12]; this review does not infer a validation method for these studies from general practice.

B. Random-Split Validation versus Spatial Generalization

The authors of [4] report AUC of 0.89±0.02 (soybean) and 0.85±0.00 (cotton) under random cross-validation, declining to 0.55–0.62 under spatial cross-validation. Because [4] addresses a different modeling task and none of the tabular classification studies performs spatial validation, this result is interpreted as a case study demonstrating the importance of testing spatial generalization, not as evidence that comparable degradation necessarily occurs across the other reviewed studies.

C. Class Balancing and Robustness

The authors of [7] report applying SMOTE and obtaining accuracy, precision, recall, F1, and ROC-AUC of 1.00 alongside sensitivity of 0.91 and specificity of 0.54 for their selected SGDC model. This review observes that the specificity value is substantially lower than the other headline metrics; it does not attribute this to SMOTE as a cause or treat it as evidence of an invalid model.

D. Dataset Provenance and Comparability

The literature exhibits inconsistent documentation of dataset provenance, sample size, and geographic origin across studies using highly similar feature schemas. The available evidence does not establish whether the datasets described are identical, overlapping, or independently constructed.

E. Temporal and External Validation

No core-verified study confirms a temporal or external validation procedure. This review distinguishes temporal coverage of a dataset (e.g., [12]'s 2001–2022 span) from temporal validation, and notes that the geographic diversity of studies across Indian states does not itself constitute external validation.

F. Reproducibility and Reporting Completeness

Information regarding code availability, dataset release, and reproducibility procedures was not confirmed in the core-verified studies examined. This is a statement about what could be confirmed, not a claim that these studies are irreproducible.

G. Implications for Robust Evaluation

Future studies would benefit from explicitly reporting validation design, testing spatial or temporal generalization where feasible, and reporting multiple complementary metrics rather than a single headline figure. This is offered as a review recommendation, not an empirically demonstrated universal requirement.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722922

Table VIII-1. Validation and generalization practices across the twelve core-verified studies.

Study	Validation Method	Spatial	Temporal	External	Class-Balancing	Tier
[1]	Hold-out 75:25; no CV	NR	NR	NR	NR	A
[2]	k-fold CV (k unspecified)	NR	NR	NR	NR	A
[3]	Train/test split; ratio NR	NR	NR	NR	NR	A
[4]	Random CV and spatial CV	Yes	NR	NR	NR	B
[5]	Hold-out 80:20 + 10-fold CV	NR	NR	NR	NR	A
[6]	Hold-out 70:30	NR	NR	NR	NR	B
[7]	Hold-out 80:20; SMOTE applied	NR	NR	NR	Yes (SMOTE)	A
[8]	NR	NR	NR	NR	NR	B
[9]	Hold-out + 5-fold CV	NR	NR	NR	NR	A
[10]	NR	NR	NR	NR	NR	D
[11]*	NR (flagged)	NR	NR	NR	NR	E
[12]	NR	NR	NR	NR	NR	B

IX. RESEARCH GAPS AND CRITICAL EVIDENCE GAPS

A. Validation and Generalization Gaps

Among the eight core-verified studies with a confirmed validation method, seven report only random hold-out or k-fold cross-validation [1], [2], [3], [5], [6], [7], [9]; only [4] additionally reports spatial cross-validation, and within its own task, performance declined substantially under geographic holdout. No reported or confirmed temporal validation procedure was identified across any of the twelve core-verified studies, and this absence of reported evidence is treated as a directly evidenced gap in what the reviewed studies report, rather than as proof that no such validation was performed. Likewise, no core-verified study reports an external validation procedure; this review therefore treats the absence of reported external and cross-region generalization testing as a directly evidenced gap in what the reviewed set reports.

B. Dataset Quality, Provenance, and Cross-Study Comparability

Five independent studies [1], [2], [5], [6], [9] report structurally similar seven-feature datasets while providing different sample-size figures and different or unstated national origins, without any providing a persistent dataset identifier. This review does not conclude that dataset duplication or leakage has occurred; the evidence supports only a documentation-inconsistency finding that limits cross-study comparability.

C. Explainability and Decision-Relevance Gaps

The authors of [1] and [3] report applying LIME; no core-verified source reports a test of whether the resulting explanations changed predictive accuracy, trust, usability, or adoption. This gap is classified as directly evidenced because it rests on a confirmed absence within a well-defined, directly examined set of studies.

D. Input-Data and Multimodal Integration Gaps

Soil moisture, electrical conductivity, and salinity are absent from all twelve core-verified studies with the partial exception of [4]'s organic carbon and CaCO3. This review treats this as a plausible but under-evidenced gap, since no reviewed study argues these variables are necessary. Multimodal integration is similarly treated as Category B. Limited real-time IoT integration within the core-verified set is treated as Category A, narrowly scoped to the twelve studies examined.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722922

E. Uncertainty and Reproducibility Gaps

Uncertainty quantification is not reported in any of the twelve core-verified studies, and no confirmation of its use was found in the sources examined; this is treated as Category B given this review's search was not designed to exhaustively cover this sub-literature. Reproducibility-related reporting was similarly not confirmed; this is not converted into a claim that the underlying studies are irreproducible.

F. Geographic and Cross-Regional Generalization

Several studies are situated within a single Indian state or district ([4], [7], [8], [9], [12]), but no core-verified study tests a single model across multiple regions within the same design. This is treated as a moderate-confidence observation given the small sample of studies reviewed.

G. Farmer-Facing Evaluation — Evidence Status: Category C

This review's search strategy was not designed to systematically identify HCI, usability, adoption, or farmer-trust literature. The present review did not specifically target this literature; therefore, the available evidence is insufficient to determine whether this represents a broader literature gap or simply a gap in this review's own search scope.

H. Synthesis of Research Gaps by Evidence Category

Category A gaps are directly evidenced by confirmed positive findings or the confirmed absence of reported evidence within the twelve studies examined. Category B gaps are plausible but insufficiently evidenced to claim with the same confidence. The single Category C item, farmer-facing evaluation, is an evidence gap in this review's own search coverage rather than a demonstrated literature gap. Table IX-1 presents each gap together with its supporting reference(s), evidence category, confidence level, and the corresponding proposed framework response developed further in Section X, consolidated into a single table to avoid duplicating the same gap list twice across two sections.

Table IX-1. Research gaps mapped to supporting evidence, confidence category, and proposed framework response (merged with the former Table X-1; no gap, reference, category, or framework mapping was removed in this consolidation).

Gap	Supporting Ref(s)	Cat.	Confidence	Proposed Framework Response (Sec. X)
Limited spatial validation	[4]; absence across [1],[2],[3],[5],[6],[7],[9]	A	High (within reviewed set)	Spatial holdout/CV — motivated by [4]
Limited temporal validation	No reported/confirmed evidence across [1]–[12]	A	High (within reviewed set)	Temporal holdout — motivated by confirmed absence
Limited external/cross-regional validation	No reported/confirmed evidence across [1]–[12]	A	High (within reviewed set)	Independent external dataset — motivated by confirmed absence
Inconsistent dataset provenance/documentation	[1], [2], [5], [6], [9]	A	High	Data governance/documentation layer
Limited evaluation of explainability (application vs. effect)	[1], [3]; qualified [10], [11]*	A	High	Explanation + evaluation + decision layer — motivated by [1],[3]
Limited cross-study comparability	[1], [2], [5], [6], [9]	A	High	Data governance/documentation layer (consequence of provenance gap)
Limited real-time IoT integration (core-verified set)	No reported/confirmed evidence across [1]–[12]	A	High (bounded to core set)	Optional real-time sensor layer

Gap	Supporting Ref(s)	Cat.	Confidence	Proposed Framework Response (Sec. X)
Limited soil-moisture/EC/salinity representation	Absence in Section III; partial exception [4]	B	Moderate	Expanded soil-data layer — evidence-informed, not an established requirement
Limited uncertainty quantification	Absence across [1]–[12]	B	Moderate	Uncertainty/calibration layer — motivated by [7]'s metric divergence
Limited reproducibility reporting	Absence of reporting across [1]–[12]	B	Moderate	Reproducible reporting layer
Limited multimodal integration	Absence across [1]–[12]	B	Moderate	Multimodal fusion layer — evidence-informed
Limited multi-region testing within a single study	[4], [7], [8], [9], [12]	B	Moderate	Multi-region external validation (Validation layer)
Farmer-facing evaluation, usability, adoption, HCI	None ([1]–[12] silent)	C	Insufficient to classify	HCI/usability/adoption evaluation — requires targeted future investigation

X. FUTURE DIRECTIONS AND PROPOSED EVIDENCE-INFORMED FRAMEWORK

The evidence reviewed in Sections III–IX indicates that AI-based crop recommendation is technically active, with demonstrated predictive applications, alongside directly evidenced limitations concerning spatial, temporal, and external validation, dataset provenance, and explainability evaluation. This section proposes a conceptual framework responding to these gaps. The framework is a synthesis produced by this review; it has not been implemented or tested, and no claim is made that it would outperform any system described in the reviewed literature.

A–D. Data, Governance, Modeling Layers

A future data-acquisition layer could incorporate soil moisture, electrical conductivity, organic carbon, and texture alongside the current N-P-K-pH-climate schema, plus IoT and remote-sensing inputs where feasible — presented as an evidence-informed direction, not a proven requirement. A proposed data-governance layer would document dataset identifiers, geographic origin, collection period, and preprocessing steps, directly motivated by the provenance inconsistency identified in Sections V and IX. A flexible modeling layer would accommodate baseline interpretable models, ensembles, and deep or hybrid architectures selected according to data characteristics, rather than prescribing a single superior algorithm.

E–H. Validation, Reliability, Explainability, Reproducibility Layers

A proposed multi-level validation strategy would sequence internal (hold-out/k-fold), spatial, temporal, and external validation — directly motivated by [4]'s finding and the confirmed absence of temporal and external validation elsewhere. A reliability layer would incorporate multiple metrics, motivated by [7]'s metric divergence. An explainability layer would separate explanation generation (as in [1], [3]) from explanation evaluation and decision-level evaluation, since the reviewed evidence demonstrates application of XAI methods but does not establish their downstream effect on trust, adoption, usability, or decision quality. A reproducibility layer would document code, data, and configuration details, responding to the confirmed absence of such reporting.

I. Deployment and Farmer-Facing Evaluation

Consistent with the Category C classification in Section IX-G, future studies could evaluate usability and adoption, but the present review cannot establish the prevalence or absence of such evaluations in the

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722922

wider literature because its search strategy did not specifically target HCI or farmer-facing decision-support research. This component is an extension requiring targeted future evidence-gathering, not a response to a confirmed field-wide gap.

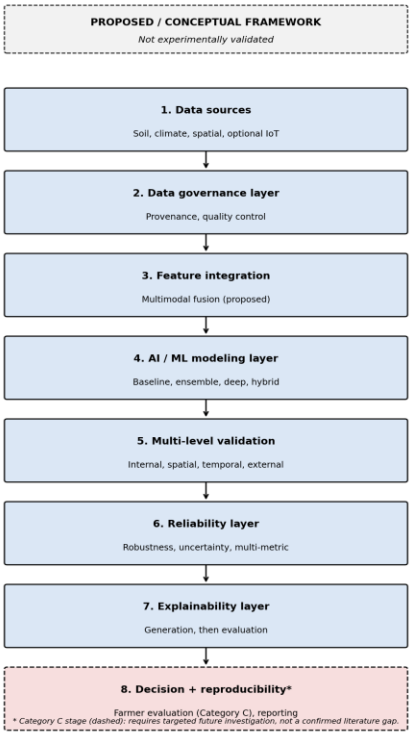

Figure X-1. Proposed evidence-informed framework for future AI-based crop recommendation systems (conceptual; not experimentally validated). The full gap-to-framework-component mapping, previously presented as a separate Table X-1, is now consolidated into Table IX-1 (Section IX-H) to avoid duplicating the same gap list across two sections; no gap, category, reference, or framework response was removed in that consolidation.

The framework proposed in this section is a conceptual synthesis derived from the evidence gaps identified in this review, mapped in detail in Table IX-1. It has not been experimentally implemented or validated, and its components should be understood as methodological recommendations for future research rather than demonstrated improvements over existing crop-recommendation systems.

XI. DISCUSSION

A. Principal Findings

The evidence reviewed indicates that AI-based crop recommendation has been implemented across a range of algorithmic approaches with several studies reporting high predictive accuracy [1], [2], [3], [5], [6], [7], [9], [12]. At the same time, these results have rarely been accompanied by validation designs capable of testing generalization beyond the training dataset. This review synthesizes these two observations as its central finding: demonstrated predictive capability alongside largely untested generalization, bounded to the twelve core-verified studies examined.

B. Predictive Performance and Algorithm Selection

This review does not rank studies by reported accuracy, because they differ in dataset composition, class count, geography, sample size, and validation design. Ensemble and boosting methods were reported as top performers in three independent studies using three different datasets [1], [5], [6], while deep-learning and graph-based models in [1] and [2] did not exceed the best ensemble results reported elsewhere. This is interpreted as evidence of competitive performance across the specific studies examined, not general superiority.

C. Validation, Generalization, and Reliability

Seven of eight core-verified studies with a confirmed validation method rely exclusively on random or k-fold validation; [4] additionally reports spatial cross-validation, finding AUC decline from 0.89/0.85 to 0.55–0.62. This is treated as a case study, not a field-wide conclusion, since none of the other reviewed models was tested this way. A second, independent line of evidence, [7]'s metric divergence (1.00 across four metrics versus 0.54 specificity), supports the same general caution: reported accuracy should be interpreted alongside validation design and metric completeness.

D. Dataset Provenance and Comparability

Documentation inconsistency across [1], [2], [5], [6], [9] is not equivalent to proven dataset duplication or leakage; none of these claims is established. What is supported is that this inconsistency limits the

interpretability of cross-study accuracy comparisons among these five sources.

E. Explainability: From Model Explanation to Decision Relevance

The reviewed evidence demonstrates application of explainability methods (LIME in [1], [3]; SHAP in the flagged [11]) but does not establish their effect on trust, usability, adoption, decision quality, or predictive accuracy. This review does not conclude that explainability is ineffective, unnecessary, or that any one method is superior.

F. Input Variables and Contextual Coverage

The narrow input schema across the benchmark-style studies leaves several agronomically recognized variables absent from the core-verified evidence. This absence is not treated as proof of agronomic irrelevance; no study tests whether including these variables would improve outcomes.

G. Relationship Between Evidence Gaps and the Proposed Framework

The framework proposed in Section X follows directly from the gaps identified in Section IX. The framework is a conceptual synthesis, not an experimentally validated solution; no claim is made that implementing it would resolve the identified gaps or improve any measured outcome.

H. Strength of Evidence and Remaining Uncertainty

Strongly established: the specific reported metrics and methods attributed to [1], [2], [3], [5], [7], and [9]. Moderately supported: the spatial-validation finding in [4] and the figures attributed to [6] (Tier B). Substantially uncertain: the prevalence of spatial-validation degradation, uncertainty quantification, reproducibility practice, and multimodal integration beyond the twelve studies, and the state of farmer-facing usability research.

I. Implications for Research and Practice

Implications developed in Section X — provenance reporting, broader validation designs, multi-metric evaluation, explicit XAI evaluation, and reproducibility practices — are presented as recommendations, not demonstrated intervention effects.

J. Limitations of the Present Review

This review's evidence base consists of twelve core-verified studies at heterogeneous verification levels: six deep-verified against full primary text ([1], [2], [3], [5], [7], [9]); [4], [8], and [12] verified at the abstract/metadata level; [6] verified at an intermediate Tier B level (authorship and methodology primary-confirmed, specific accuracy figures cross-source verified); [10] confirmed only bibliographically; and [11] flagged, with authorship unconfirmed despite two independent verification attempts. This review's search strategy was not designed to identify HCI or farmer-adoption literature, and this review was unable to establish whether the datasets in [1], [2], [5], [6], [9] are identical, overlapping, or independently constructed.

XII. CONCLUSION

A. Overall Conclusion

This review set out to critically synthesize what soil and climate data and AI methods have been used for crop recommendation, how reliable the reported results are, and what methodological gaps remain. Based on twelve core-verified studies, this review finds that AI-based crop recommendation has been implemented across classical machine learning, ensemble, and deep-learning approaches, with several studies reporting high predictive accuracy, while these results have rarely been accompanied by validation designs capable of testing generalization beyond the training dataset.

B. Principal Methodological Findings

Validation practice across the eight core-verified studies with a confirmed method is dominated by random or k-fold cross-validation; only [4] additionally reports spatial cross-validation, finding substantial performance decline within its own species-distribution-modeling task — a case study, not a field-wide conclusion. The literature exhibits inconsistent documentation of dataset provenance, sample size, and geographic origin across studies using highly similar feature schemas [1], [2], [5], [6], [9]. Explainability methods were applied in two deep-verified studies [1], [3], but their effect on trust, usability, or adoption was not tested. No core-verified study confirmed a temporal or external validation procedure, and several agronomically recognized soil

and climate variables were absent from the input schemas examined.

C. Evidence-Informed Research Priorities

Future research would benefit from standardized dataset-provenance reporting, validation designs incorporating spatial and temporal testing, multi-metric evaluation reporting, explicit evaluation of explainability's downstream effects, and more complete reproducibility reporting. Farmer-facing usability and adoption evaluation is identified as an area requiring targeted future investigation rather than a confirmed gap.

D. Final Epistemic Boundary

The findings and priorities presented in this review are bounded to the twelve core-verified studies examined, at their respective verification tiers, and should not be read as characterizing AI-based crop recommendation research as a whole. The conceptual framework proposed in Section X has not been implemented or experimentally validated, and no claim is made that it would improve predictive accuracy, generalization, trust, or adoption relative to the systems described in the reviewed literature.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

REFERENCES

[1] S. Shastri, S. Kumar, V. Mansotra, and R. Salgotra, "Advancing crop recommendation system with supervised machine learning and explainable artificial intelligence," *Scientific Reports*, vol. 15, art. 25498, 2025. Crossref

[2] P. Ayesha Barvin and T. Sampradeepraj, "Crop recommendation systems based on soil and environmental factors using graph convolution neural network: A systematic literature review," *Engineering Proceedings*, vol. 58, no. 1, art. 97, 2023. Crossref

[3] M. Y. Shams, S. A. Gamel, and F. M. Talaat, "Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making," *Neural Computing and Applications*, vol. 36, no. 11, pp. 5695–5714, 2024. Crossref

[4] S. Moharana, S. R. Naitam, D. O. Shirale, S. K. Adamala, N. Dash, C. Amrutha, A. Raghuvanshi, S. Karthikeyan, D. R. Biswas, and A. Patil, "Modeling climate-resilient crop suitability in central India using machine learning and species distribution modeling approaches," *Theoretical and Applied Climatology*, vol. 156, art. 604, 2025.

A. Afzal, S. Amjad, M. Raza, M. Munir, F. Villar, and A. Ashraf, "Incorporating soil information with machine learning for crop recommendation to improve agricultural output," *Scientific Reports*, vol. 15, art. 8560, 2025.

B. Dey, J. Ferdous, and R. Ahmed, "Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables," *Heliyon*, vol. 10, no. 3, art. e25112, 2024. Crossref [Tier B — authorship, affiliation, and methodology confirmed via primary/near-primary sources; specific accuracy figures remain cross-source verified]

[5] M. K. Senapaty, A. Ray, and N. Padhy, "A decision support system for crop recommendation using machine learning classification algorithms," *Agriculture*, vol. 14, no. 8, art. 1256, 2024. Crossref

[6] S. Rani, A. K. Mishra, A. Kataria, S. Mallik, and H. Qin, "Machine learning-based optimal crop selection system in smart agriculture," *Scientific Reports*, vol. 13, art. 15997, 2023. Crossref

[7] G. Singh and S. Sharma, "Enhancing precision agriculture through cloud based transformative crop recommendation model," *Scientific Reports*, vol. 15, art. 9138, 2025. Crossref

[8] Y. Akkem, S. K. Biswas, and A. Varanasi, "Role of explainable AI in crop recommendation technique of smart farming," *International Journal of Intelligent Systems and Applications*, vol. 17, no. 1, pp. 31–52, 2025. Crossref

[9] Authors not independently confirmed, "Interpretable deep learning models for independent fertilizer and crop recommendation," *Scientific Reports*, 2025.

[FLAGGED — authorship unconfirmed after two independent verification attempts; reported descriptively only and not used to support any conclusion beyond this qualified instance]

[10] Y. Mahale, N. Khan, K. Kulkarni, S. A. Wagle, P. Pareek, K. Kotecha, T. Choudhury, and A. Sharma, "Crop recommendation and forecasting system for Maharashtra using machine learning with LSTM: a novel expectation-maximization technique," *Discover Sustainability*, vol. 5, no. 1, pp. 1–23, 2024.