JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Explainable Ensemble Learning for Crop Yield Prediction: A Review of Methods, Validation Practices, and Research Challenges

[1]Dr. SYED GILANI PASHA **[2]Dr.RAVI CHINKERA**

[1]Professor [2]Associate Professor
[1,2]Department of Computer Science & Engineering
[1,2]SECAB Institute of Engineering & Technology, Vijayapura, Karnataka, India

Abstract : **Crop yield prediction supports agricultural planning, resource management, and food security, but remains challenging because of complex interactions among soil properties, climatic conditions, crop characteristics, and management practices. Ensemble learning offers flexible approaches for modelling these relationships, while explainable artificial intelligence provides methods for examining the factors influencing predictions. This review examines bagging, Random Forest, Gradient Boosting, and stacking approaches for crop yield prediction, alongside explainability techniques such as feature importance, partial dependence plots, and SHapley Additive exPlanations. Particular attention is given to dataset characteristics, predictor availability, preprocessing, and validation practices, including temporal and geographical evaluation. The discussion distinguishes predictive performance from generalization and considers how data leakage, inconsistent evaluation protocols, and correlated predictors can affect the interpretation of findings. Key research challenges include limited representative datasets, missing observations, regional variability, changing climatic conditions, and the difficulty of translating model explanations into actionable agricultural guidance. Future research priorities include reproducible benchmarking, uncertainty estimation, validation across growing conditions, and the integration of agricultural knowledge into model development. The review emphasizes that useful crop yield prediction requires a balanced assessment of accuracy, interpretability, reliability, and operational relevance.**

IndexTerms **- Crop yield prediction; ensemble learning; explainable artificial intelligence; model validation; SHAP; agricultural forecasting**

I. INTRODUCTION

Crop yield prediction supports agricultural planning by providing estimates that can inform resource allocation, production monitoring, storage, and distribution[1-2]. Developing reliable predictions is challenging because yield depends on interacting factors, including weather conditions, soil properties, crop characteristics, and management practices[3-4]. These relationships vary across locations and growing seasons, making it necessary to evaluate predictive models under the conditions in which they will be used. The reviewed studies address this complexity using combinations of environmental, management, and historical yield information[5-6]. Statistical approaches such as linear regression provide useful baselines for examining relationships between agricultural predictors and yield[7-8]. However, simple linear formulations may not adequately represent nonlinear responses and interactions. Ensemble learning offers alternative approaches by combining multiple estimators. Jeong et al. (2016), for example, found that Random Forest outperformed Multiple Linear Regression across their evaluated wheat, maize, and potato datasets. Their findings also highlighted limitations when predicting extreme outcomes or responses beyond the training range.

Ensemble methods differ in how they create and combine component models. Bagging aggregates estimators trained on bootstrap samples, Random Forest introduces additional predictor randomization, boosting builds estimators sequentially, and stacking learns a combination of base-model predictions. Agricultural research has also explored weighted ensembles and ensembles of neural networks. These approaches can improve prediction, but their benefits depend on the available data and the diversity of component errors. Shahhosseini et al. (2020) demonstrated that weighted combinations could outperform stacking in their corn-forecasting setting, while their subsequent CNN–DNN study extended ensemble construction to neural architectures.

Predictive performance alone provides an incomplete basis for assessing agricultural models. Researchers and intended users also need to understand which variables influence predictions and whether the model relies on plausible relationships. Explainable artificial intelligence provides tools for examining these questions through global descriptions of model behaviour and local explanations of individual predictions. Relevant approaches include feature importance, partial dependence plots, accumulated local effects, and SHapley Additive exPlanations. Srivastava et al. (2022) illustrated the application of SHAP to winter wheat prediction by examining environmental and phenological contributions at both global and individual-observation levels.

Nevertheless, explanations require careful interpretation. Correlated predictors can obscure individual importance values, and partial dependence plots may evaluate combinations poorly represented in the observed data. ALE was developed to reduce this

extrapolation problem, while SHAP provides feature attribution relative to a reference. These methods describe model behaviour; their outputs do not independently establish causal relationships or justify changes to agricultural management. Consequently, explanation quality should be considered alongside predictive validity rather than treated as proof of model reliability.

Validation practices introduce another important distinction between estimating yield and forecasting it under operational constraints. A model evaluated using complete growing-season information addresses a different task from one issuing an early-season forecast. Similarly, performance on familiar locations does not establish transferability to unseen regions. Shahhosseini et al. (2020) addressed temporal structure through sequential validation, later-year testing, and experiments with partial-season weather information. Their approach demonstrates why prediction timing and information availability must be considered when interpreting reported performance.

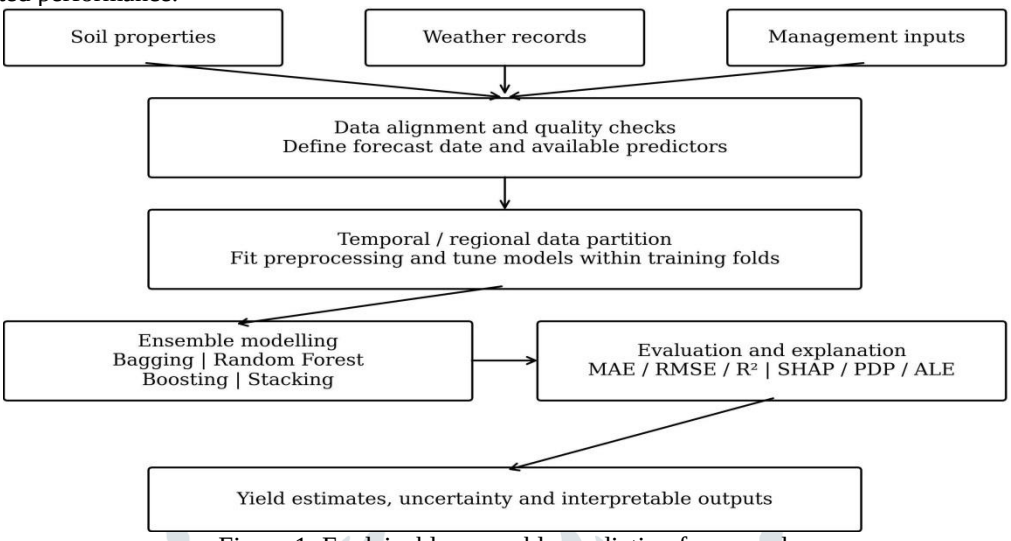


Figure 1: Explainable ensemble prediction framework

Against this background, this narrative review examines explainable ensemble learning for crop yield prediction, focusing on methods, validation practices, and research challenges. It compares selected published studies according to their datasets, predictors, algorithms, interpretation techniques, and evaluation conditions. The review also discusses data representativeness, generalization, information leakage, explanation stability, uncertainty, and practical usability. Its purpose is to identify the conditions under which ensemble predictions and their explanations can provide credible support for agricultural decision-making, while outlining priorities for reproducible and operationally relevant research.

II. Review of Related Works

Research on crop yield prediction has progressed from comparisons between linear regression and tree ensembles toward hybrid architectures and explainable modelling. The selected studies below illustrate this development through differences in datasets, predictors, algorithms, evaluation practices, and interpretation methods. Their performance values describe their respective experimental settings and should not be treated as a direct ranking across studies.

2.1 Ensemble Methods and Predictive Performance

Jeong et al. (2016) demonstrated the value of Random Forest as an alternative to linear regression across several crop datasets. Their findings support the ability of tree ensembles to represent relationships that a linear formulation may miss. However, their caution regarding extreme values highlights an important distinction between fitting patterns within the observed data range and extrapolating to unfamiliar conditions

Shahhosseini, Hu, and Archontoulis (2020) extended this comparison by evaluating different ways of combining models. Weighted and average ensembles produced the most precise forecasts, while stacked LASSO produced the least biased predictions in their evaluation. This distinction shows why model selection should consider multiple criteria: the model with the smallest overall prediction error may differ from the model with the smallest average bias.

The subsequent CNN–DNN ensemble study broadened ensemble construction to include neural architectures. Its comparison of homogeneous and heterogeneous ensembles showed that diversity could be introduced through training samples as well as model configurations. The reported advantage of the homogeneous ensemble demonstrates the potential of bagging neural models under the study's data and evaluation conditions.

2.2 Explainability and Agricultural Interpretation

The reviewed studies employ different levels of interpretation. Jeong et al. used variable importance and partial dependence plots, while Shahhosseini et al. developed partial-dependence-based analysis for their weighted ensemble. These approaches examine overall model behaviour and the relationships between predictors and estimated yield.

Srivastava et al. used SHAP to examine both overall feature importance and contributions to individual wheat-yield predictions. Gupta et al. incorporated SHAP into feature selection for Indian wheat prediction. Together, these studies illustrate two distinct uses of explainability: interpreting a fitted model and supporting the selection of its inputs.

2.3 Validation Practices and Remaining Gaps

Validation design is a central consideration when interpreting these findings. Shahhosseini et al. (2020) explicitly separated later test years from model development and used a walk-forward procedure for tuning. Their evaluation also considered partial-season weather availability, connecting model assessment to the information available when a forecast would actually be issued.

The comparison suggests that a stronger review should assess each model alongside its prediction date, geographical coverage, input requirements, and validation strategy. Differences in crops, yield variability, units, and test conditions prevent direct comparisons between the numerical errors in Table 1. Further research should therefore examine whether predictive

improvements remain consistent across future seasons and unseen regions, and whether model explanations remain useful under those changing conditions.

Table 1: Comparison of selected studies on ensemble learning and explainable crop yield prediction

Study	**Dataset and coverage**	**Main predictors**	**Algorithms and interpretation**	**Findings and relevance**
Jeong et al. (2016)	Global gridded wheat data, US county-level maize records over 30 years, and regional potato and maize silage datasets	Climate and biophysical variables, including precipitation, temperature, fertilizer inputs, and location or year, depending on the dataset	Random Forest compared with Multiple Linear Regression; variable importance and partial dependence plots	Random Forest outperformed the linear benchmarks across the evaluated cases. For global wheat, RMSE was 0.32 tonnes/ha for Random Forest and 1.32 tonnes/ha for the linear model. The authors cautioned about predictions at extremes and beyond the training range.
Shahhosseini, Hu, and Archontoulis (2020)	5,342 observations from 293 **counties** in Illinois, Indiana, and Iowa, covering 2000–2018	Weekly weather, soil characteristics, planting progress, plant population, and yield-trend features	Weighted, average, and stacked ensembles; partial-dependence-based feature analysis	Optimized weighted and average ensembles achieved 9.5% relative RMSE in the reported evaluation. The study used sequential validation and reserved 2016–2018 for testing, providing a useful example of forecasting-oriented evaluation.
Shahhosseini et al. (2021)	County-level corn data from 12 US Corn Belt states**,** covering 1980–2019	Environmental and management variables, soil and weather inputs, and historical yields	Homogeneous and heterogeneous CNN–DNN ensembles constructed through bagging, weighted aggregation, and stacking	The homogeneous ensemble achieved a 2019 RMSE of 866 kg/ha, equivalent to 8.5% relative RMSE, and explained approximately 77% of spatiotemporal yield variation. The work extends ensemble modelling beyond conventional tree-based methods.
Srivastava et al. (2022)	Winter wheat data from 271 German counties**,** covering 1999–2019	Weekly weather, soil water characteristics, bulk density, and sowing, flowering, and harvest dates	CNN compared with eight baseline models, including Random Forest and XGBoost; SHAP interpretation	The CNN achieved 7–14% lower RMSE than the best-performing baseline across the evaluated test data. SHAP analysis examined global feature importance and individual predictions. Although the proposed model was not an ensemble, the study provides a relevant benchmark for explainability and ensemble comparisons.
Gupta et al. (2026)	Indian wheat data from CY-Bench	NDVI, meteorological variables, and soil moisture	Voting, stacking, and boosting ensembles combined with SHAP-based feature selection	The authors reported that the Boost m1 ensemble outperformed the alternatives and that SHAP-based selection improved prediction and interpretation. The publisher abstract does not provide numerical performance values, so none are reproduced here.

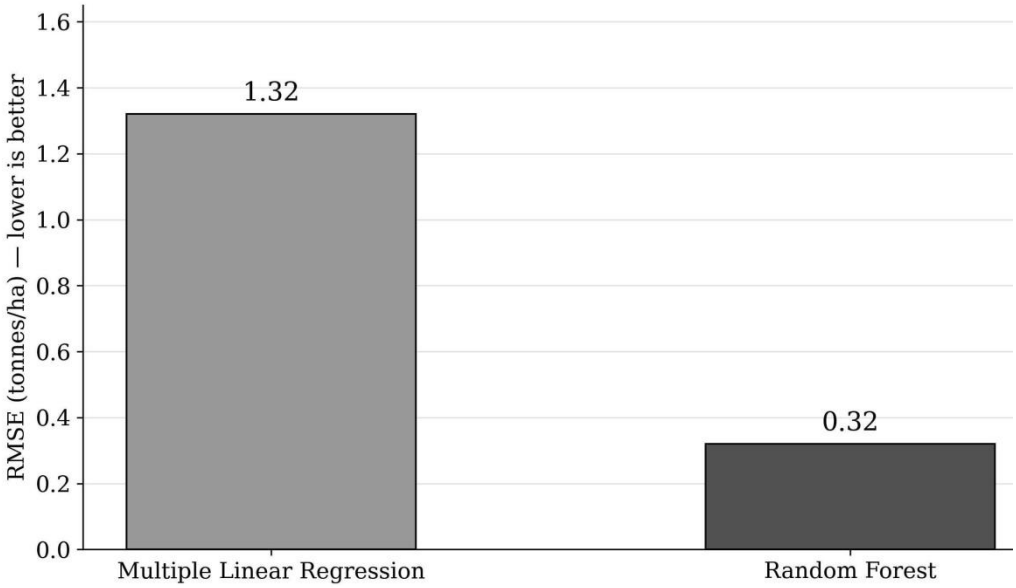


Figure 2: Published RMSE comparison: Random Forest versus Linear Regression

Figure 2 uses published results from Jeong et al. (2016); the other figures are conceptual illustrations.

III. ENSEMBLE LEARNING METHODS FOR CROP YIELD PREDICTION

Ensemble learning combines predictions from multiple models to estimate crop yield. Its effectiveness depends on the quality of the individual models and the diversity of their errors. Agricultural datasets contain interacting soil, climatic, and management variables, making ensemble approaches useful for investigating complex yield relationships. The reviewed literature includes tree ensembles, combinations of different regression algorithms, and ensembles of neural networks.

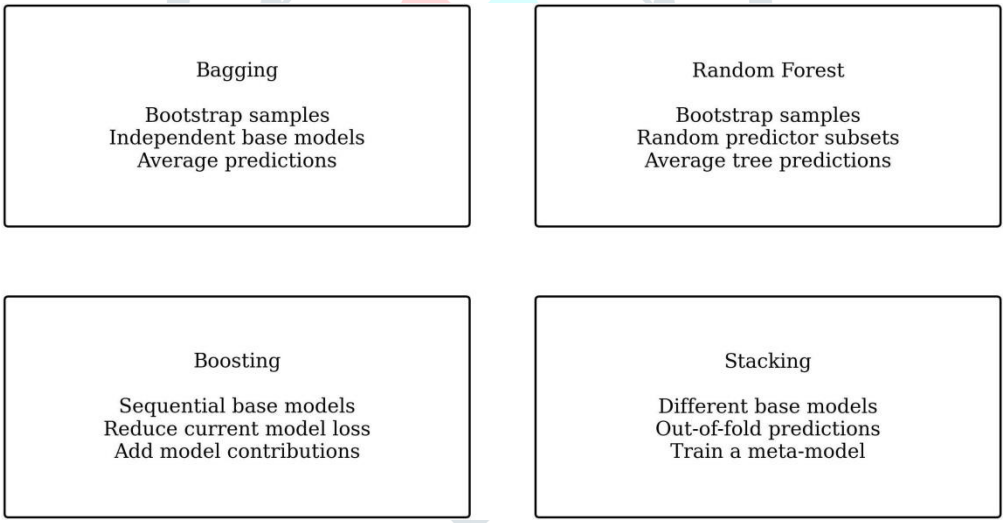


Figure 3: Comparison of ensemble learning methods

3.1 Bagging Regression

Bootstrap aggregating, commonly called bagging, trains multiple base estimators on samples drawn with replacement from the training dataset. For regression, their predictions are usually averaged:

$$\hat{y}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x)$$

where B denotes the number of estimators and $\hat{f}_b(x)$ represents the prediction of the b-th estimator. By averaging models that respond differently to variations in training data, bagging can reduce prediction variance. Its benefits are greatest when individual estimators are unstable and their errors are not strongly correlated.

In crop yield prediction, bagging can combine decision trees that capture different relationships among weather, soil, and management inputs. The principle also extends to neural models. Shahhosseini et al. (2021) used bagged training datasets to introduce diversity among CNN–DNN models with the same architecture. Their findings demonstrate that ensemble diversity can arise from differences in training samples, rather than requiring different algorithm types.

3.2 Random Forest Regression

Random Forest extends tree bagging by introducing random selection of candidate predictors during tree construction. This additional randomness helps reduce similarity among the trees, whose predictions are averaged to produce the final yield estimate.

The approach can represent nonlinear relationships and interactions without requiring these relationships to be specified in advance.

Jeong et al. (2016) evaluated Random Forest against Multiple Linear Regression for wheat, maize, and potato yield prediction. Random Forest performed better across their evaluated cases. The study also used variable importance and partial dependence plots to examine model behaviour. However, the authors identified limitations when predicting extreme outcomes or responses outside the training range. These findings indicate that strong test performance within represented conditions does not establish reliable extrapolation to unfamiliar agricultural environments.

3.3 Boosting Methods

Boosting constructs an ensemble sequentially, with successive estimators addressing weaknesses in the current model. In gradient boosting regression, a new estimator is fitted to a direction that reduces the selected loss function. A simplified update is:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

where $F_{m-1}(x)$ is the existing ensemble, $h_m(x)$ is the newly fitted estimator, and η is the learning rate. XGBoost and LightGBM are implementations of gradient-boosted decision trees included in the reviewed corn-yield research.

For crop yield prediction, boosting provides a means of progressively modelling relationships that earlier estimators have not captured adequately. Its performance depends on settings such as tree complexity, learning rate, and the number of boosting iterations. These choices should be evaluated under validation conditions consistent with the intended forecasting task. In the study by Shahhosseini et al. (2020), boosted models were evaluated alongside other learners and incorporated into broader ensemble comparisons, rather than assumed to be universally superior.

3.4 Averaging and Weighted Ensembles

Averaging ensembles combine predictions from several models using equal weights. Weighted ensembles assign different contributions to individual models:

$$\hat{y}(x) = \sum_{j=1}^{M} w_j \hat{f}_j(x), \qquad \sum_{j=1}^{M} w_j = 1$$

where M is the number of component models and w_j is the weight assigned to model j. A weighted average commonly uses non-negative weights estimated from development data.

Shahhosseini et al. (2020) found that optimized weighted and average ensembles produced the most precise forecasts in their corn-yield evaluation. Their findings support examining whether combining complementary models improves predictions. They also show why ensemble selection should consider the intended objective, since the model with the lowest prediction error differed from the model with the lowest bias.

3.5 Stacking

Stacking uses a meta-model to learn how to combine predictions from multiple base models. Unlike a fixed averaging rule, it estimates the relationship between the component predictions and observed yield. Base models may include linear regression, Random Forest, and boosting algorithms, allowing the ensemble to combine different modelling assumptions.

The meta-model must be trained using predictions generated for observations excluded from the corresponding base-model training. For agricultural time series, these predictions should also respect chronological order. Shahhosseini et al. (2020) used a blocked sequential procedure and found that stacking did not consistently outperform weighted ensembles. Thus, adding a learned combination layer does not automatically improve forecasting performance.

3.6 Deep Learning Ensembles

Deep learning ensembles combine multiple neural networks, potentially using different training samples, model settings, or architectures. In crop yield modelling, these networks can process environmental sequences and other agricultural inputs before their predictions are aggregated.

Shahhosseini et al. (2021) examined homogeneous and heterogeneous CNN–DNN ensembles for county-level corn prediction. Their architecture used convolutional components to extract features from weather and soil inputs, followed by dense layers for prediction. The study illustrates how ensemble principles can be applied to neural architectures as well as conventional regression models.

3.7 Implications for Method Selection

The reviewed studies indicate that ensemble design should be assessed alongside input availability, prediction timing, and validation strategy. Bagging and Random Forest emphasize aggregation of diverse estimators, boosting develops models sequentially, and stacking learns how to combine model outputs. Their relative value must be established within a common evaluation setting.

For the present review, a useful comparison therefore considers predictive error, bias, computational requirements, interpretability, and performance across unseen seasons or regions. This assessment follows the distinction in the reviewed research between accurate yield estimation under available data and forecasting under the information constraints faced during a growing season.

IV. Explainable AI Approaches

Explainable artificial intelligence (XAI) provides methods for examining how predictive models use input variables to produce their outputs. In crop yield prediction, explanations can help researchers investigate whether a model relies on agriculturally plausible relationships and identify why predictions differ across observations. Explanations may be **global**, describing overall model behaviour, or **local**, describing a particular prediction. SHAP provides a framework for feature attribution, while LIME constructs interpretable approximations around individual predictions.

4.1 Feature Importance

Feature importance summarizes the contribution or predictive relevance of input variables within a fitted model. For crop yield prediction, it can help identify whether the model relies primarily on weather, soil characteristics, management inputs, or

historical yield patterns. However, an importance ranking alone does not explain whether higher values of a predictor increase or decrease estimated yield.

Tree ensembles commonly provide importance measures based on reductions in the splitting criterion. These measures require caution because they can favour variables with many possible split points and may reflect patterns that do not generalize. **Permutation importance** offers an alternative by measuring the decline in model performance after a predictor is shuffled. Evaluation on held-out observations helps assess reliance on that predictor beyond the training data. Correlated predictors can nevertheless make individual importance values misleading because one variable may substitute for another.

4.2 SHapley Additive exPlanations

SHapley Additive exPlanations (SHAP), introduced by Lundberg and Lee, assigns feature contributions to a prediction relative to a reference value. For an additive explanation of a regression output, the relationship can be expressed as:

$$\hat{f}(x) = \phi_0 + \sum_{j=1}^{p} \phi_j$$

where ϕ_0 is the reference prediction and ϕ_j represents the contribution attributed to feature j. A positive contribution raises the prediction relative to the reference, while a negative contribution lowers it. Averaging absolute contributions across observations provides a summary of global importance. The reference data and assumptions used to handle unavailable features influence the resulting explanation

In agricultural modelling, SHAP can examine the contributions of rainfall, temperature, soil properties, and phenological variables to estimated yield. Srivastava et al. (2022) applied SHAP to winter wheat prediction in Germany, using global summaries and local force plots to interpret environmental and phenological inputs. Their analysis illustrates how explanations can examine both overall model behaviour and differences among individual yield estimates. These are explanations of model predictions and should not be interpreted as causal effects on crop production.

4.3 Partial Dependence Plots

Partial dependence plots (PDPs) summarize how average model predictions change as a selected predictor varies. For example, a rainfall PDP can display the model's average predicted yield across a range of rainfall values. Two-variable plots can also help examine interactions, such as the combined relationship of temperature and precipitation with predicted yield.

PDPs require particular care when predictors are correlated. Varying one variable while retaining observed values of the others may generate combinations rarely represented in the dataset. A rainfall–soil-moisture combination created during this calculation, for instance, may be unrealistic. Consequently, the resulting curve can reflect model extrapolation rather than behaviour supported by observed agricultural conditions. Apley and Zhu examined this limitation when developing accumulated local effects plots.

4.4 Accumulated Local Effects

Accumulated local effects (ALE) plots examine changes in predictions over small intervals of a predictor and accumulate these changes across its observed range. This approach reduces the need to evaluate unrealistic combinations of correlated inputs. ALE is therefore relevant to agricultural datasets in which weather and soil variables are strongly related.

Apley and Zhu proposed ALE as an alternative to partial dependence for such settings. Nevertheless, ALE remains a description of a fitted model, and estimates in sparsely populated intervals require caution. It complements other explanation methods by showing the shape of model responses rather than providing only an importance ranking.

4.5 Local Interpretable Model-Agnostic Explanations

Local Interpretable Model-Agnostic Explanations (LIME) explains an individual prediction by fitting a simpler surrogate model around the observation of interest. The surrogate approximates the complex model within that neighbourhood, allowing its local behaviour to be presented in an interpretable form. Ribeiro, Singh, and Guestrin introduced this approach to explain predictions without requiring access to the internal structure of the original model.

For crop yield regression, a local surrogate could describe which inputs influence a selected estimate. Its interpretation is limited to the neighbourhood used for fitting; it should not be treated as a global description of the ensemble. As a methodological recommendation, agricultural applications should check whether generated neighbouring observations are plausible and whether the explanation remains similar under repeated runs.

4.6 Comparison and Interpretation

The table 2 shows that explainable AI approaches provide complementary insights into crop yield prediction models. Feature importance identifies the variables on which a model relies, while permutation importance examines how shuffling a predictor affects predictive performance. These methods help identify influential inputs but do not directly explain the direction of their relationship with estimated yield.

SHAP explains how individual inputs contribute to predictions relative to a reference value, supporting both overall and observation-specific interpretation. PDP illustrates average prediction responses as a variable changes, whereas ALE examines accumulated local changes with less extrapolation when predictors are correlated. LIME provides a simpler approximation of model behaviour around an individual prediction.

The choice of method should depend on the explanation required, predictor relationships, and data coverage. Using complementary methods can provide a broader understanding of model behaviour. However, these explanations describe how the model generates predictions; they do not establish that changing an agricultural input will cause a corresponding change in actual yield.

Table 2: Explainable AI methods relevant to crop yield prediction

Approach	Main purpose	Agricultural interpretation	Principal limitation
Feature importance	Rank influential predictors	Identify variables on which the model relies	Rankings do not show response direction; some measures are biased
Permutation importance	Measure performance loss after shuffling a	Assess predictive reliance on held-out	Correlated variables can mask individual

	predictor	data	importance
SHAP	Attribute predictions to individual inputs	Explain overall patterns and selected yield estimates	Results depend on reference data and feature-dependence assumptions
PDP	Display average prediction responses	Examine nonlinear response patterns	Correlated inputs can create unrealistic combinations
ALE	Accumulate local prediction changes	Examine responses with less extrapolation under correlated inputs	Sparse intervals can produce uncertain estimates
LIME	Approximate a prediction locally	Explain an individual estimate using a simpler model	Interpretation depends on the selected neighbourhood

For this review, explanation quality should be assessed alongside predictive performance. Useful evaluation questions include whether explanations remain consistent across model refits, whether they change across seasons or regions, and whether their interpretation agrees with agricultural knowledge. Agreement between methods can strengthen confidence in an interpretation, but it does not establish causality or validate a management recommendation. Recommendations concerning irrigation or fertilizer use require evidence beyond feature attribution.

V. Discussion

The reviewed studies indicate that ensemble learning can improve crop yield prediction by representing relationships among environmental conditions, soil characteristics, and agricultural management variables. However, the evidence does not establish a universally superior algorithm. Jeong et al. (2016) found that Random Forest outperformed Multiple Linear Regression across the evaluated crop datasets, while Shahhosseini et al. (2020) demonstrated advantages for weighted and average ensembles in corn forecasting. These findings support ensemble modelling as a useful approach, but their implications remain specific to the datasets, prediction tasks, and evaluation conditions examined. j

An important distinction emerges between model complexity and effective ensemble construction. Shahhosseini et al. (2020) found that stacking did not consistently outperform simpler weighted combinations. Their results also distinguished prediction precision from bias, with different ensemble strategies performing best under different criteria. Similarly, the CNN–DNN ensemble study demonstrated the value of introducing diversity through bagged training samples. Together, these studies suggest that complementary prediction errors and appropriate combination procedures can matter as much as the complexity of individual learners.

Dataset composition further shapes the interpretation of model performance. The reviewed investigations differ in crop type, geographical coverage, observation period, spatial resolution, and predictor availability. The US corn studies combined weather, soil, management, and historical information, whereas the German winter wheat study incorporated environmental and phenological variables. Consequently, differences in reported errors cannot be attributed solely to algorithm selection. A lower error in one publication does not demonstrate superiority over a model evaluated on another crop or under different growing conditions. The evidence therefore supports comparisons within shared experimental settings rather than direct rankings across publications.

Validation practices are particularly important when distinguishing yield estimation from operational forecasting. Shahhosseini et al. (2020) reserved later years for testing and used sequential validation during model development. Their examination of partial-season weather information also addressed the practical question of when a prediction can be issued. This approach provides a stronger basis for assessing future-season forecasting than an evaluation that mixes observations across years without preserving their temporal structure. Nevertheless, successful prediction in later years within familiar locations does not, by itself, establish transferability to unseen regions.

Explainability adds another dimension to model assessment. Jeong et al. (2016) used variable importance and partial dependence plots, while Srivastava et al. (2022) applied SHAP to examine global patterns and individual predictions. These approaches help investigate which inputs influence model outputs and how their contributions differ across observations. Their value extends beyond presenting an importance ranking: they can guide scrutiny of whether predictions depend on plausible agricultural information. However, explanations describe the fitted model's behaviour and do not establish the effects of interventions such as increasing irrigation or fertilizer application

The reliability of explanations also requires critical examination. Correlated agricultural predictors can complicate both importance rankings and response plots. Permutation importance may understate the contribution of a variable when another predictor provides similar information, while partial dependence can evaluate combinations that are poorly represented in the data. ALE offers an alternative that reduces this extrapolation problem, and SHAP provides additive feature attribution relative to a reference. These methods answer different questions, so their outputs should be interpreted according to their assumptions rather than treated as interchangeable evidence.

The synthesis identifies three priorities for assessing explainable ensemble systems: predictive performance under realistic information constraints, generalization across growing conditions, and the stability and usefulness of explanations. Future comparisons should document forecast timing, input availability, preprocessing, validation partitions, and subgroup errors. Explanations should also be examined across seasons, regions, and model refits. This broader assessment would help determine whether an apparently accurate model can provide consistent and interpretable support for agricultural decisions, while keeping conclusions within the conditions actually evaluated.

VI. Research Gaps and Challenges

6.1 Limited Representative Data

A major challenge is the availability of datasets that adequately represent different crops, soil types, management practices, and growing conditions. The reviewed studies provide evidence from particular agricultural settings, including US corn production and German winter wheat cultivation. Their findings cannot automatically be extended to other regions or crops. Shahhosseini et al. (2020) also acknowledged that some inputs used in their framework may be less readily available elsewhere. An important research gap is therefore the evaluation of models under both diverse agricultural conditions and limited input availability.

6.2 Data Quality and Heterogeneity

Agricultural prediction often requires integrating weather measurements, soil information, management records, and yield statistics collected at different spatial and temporal resolutions. Shahhosseini et al. (2020), for example, combined soil, weather, and management information with differing resolutions, while Srivastava et al. (2022) aggregated environmental and phenological information to county-level observations. These procedures create methodological choices concerning alignment, aggregation, and the representation of local conditions. A remaining challenge is to quantify how these choices, together with missing or uncertain measurements, affect predictions and explanations.

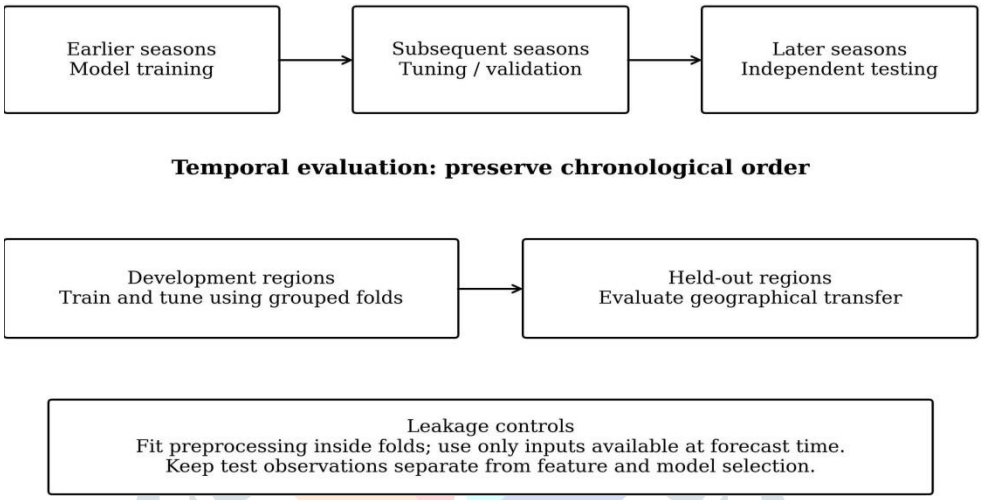


Figure 4: Temporal and geographical validation

6.3 Generalization Across Regions and Seasons

Models that perform well within a familiar dataset may be less reliable under different environmental or management conditions. Jeong et al. (2016) explicitly identified limitations when Random Forest predictions extend toward extremes or beyond the training range. Although later-year testing provides useful evidence of temporal generalization, it does not establish geographical transferability. Further research should distinguish prediction in familiar locations, forecasting future seasons, and prediction in previously unseen regions, using separate validation designs for each task.

6.4 Validation Design and Information Leakage

Validation must reflect the intended application. Agricultural observations may share information through location, season, or repeated measurements, making an unrestricted random split unsuitable for some forecasting tasks. Shahhosseini et al. (2020) addressed temporal structure through sequential validation and later-year testing. A broader methodological challenge is ensuring that preprocessing, feature selection, hyperparameter tuning, and ensemble construction use only the permitted development data. For stacking, the predictions used to train the meta-model must also be generated without exposing the corresponding base models to those observations.

6.5 Forecast Timing and Predictor Availability

The usefulness of a yield forecast depends on when it becomes available. A model using complete growing-season information addresses a different task from one predicting yield before planting or during early crop development. The partial-season experiments of Shahhosseini et al. (2020) demonstrate the importance of this distinction. Future evaluations should explicitly define the forecast date and identify which inputs are available at that time. Otherwise, strong predictive performance may overstate the model's usefulness for timely agricultural decisions.

6.6 Comparability and Reproducibility

The selected studies differ in crop coverage, geographical scale, predictors, and evaluation periods. Their numerical findings therefore cannot establish a common ranking of ensemble methods. Even within one study, the preferred model can depend on whether the objective is to minimize error or bias. A research priority is reproducible benchmarking with shared datasets, identical validation partitions, clearly defined metrics, and comparable tuning procedures. Such comparisons should retain simple baselines so that the benefits of additional model complexity can be assessed.

6.7 Reliability of Explanations

Feature importance and explanation plots do not provide an unambiguous account of agricultural processes. Correlated predictors can obscure individual permutation importance, and partial dependence plots may evaluate unrealistic combinations of inputs. ALE reduces this extrapolation problem, while SHAP attributes predictions relative to a specified reference. The methodological gap is to evaluate explanation stability across training samples, model configurations, seasons, and regions, rather than relying on a single ranking or plot. s

6.8 Causal Interpretation and Actionability

Explaining a prediction is different from establishing the consequences of changing an agricultural practice. For example, a positive contribution attributed to fertilizer in a fitted model does not demonstrate that increasing fertilizer would improve yield for a particular farm. The reviewed explanation methods characterize model behaviour; translating their outputs into recommendations requires additional agronomic and causal evidence. Future work should assess whether explanations are understandable to intended users and whether proposed actions are feasible under local resource constraints.

6.9 Uncertainty and Extreme Conditions

The reviewed comparisons emphasize point-prediction performance, including RMSE and bias. These measures do not, by themselves, communicate the uncertainty associated with an individual forecast. This is particularly relevant when a model encounters conditions poorly represented during training. A priority arising from this review is to evaluate prediction intervals, their coverage, and their usefulness across ordinary and extreme seasons. Such evaluation should complement, rather than replace, the assessment of average prediction error.

VII. Future Research Directions

7.1 Representative Agricultural Datasets

Future research should prioritize datasets covering diverse crops, geographical regions, soil conditions, and management practices. Greater representation of smallholder farms, rainfed agriculture, and extreme growing conditions would support more informative assessments of model transferability. Dataset documentation should specify collection procedures, measurement units, missing values, and spatial and temporal coverage. Where data sharing is restricted, common reporting standards and evaluation protocols could still improve comparisons across studies.

7.2 Realistic Temporal and Geographical Validation

Validation should match the intended prediction task. Future-season forecasting requires chronological evaluation, while geographical transfer requires testing on locations excluded from model development. Combining these approaches would provide a stronger assessment of performance under unfamiliar conditions. The sequential validation used by Shahhosseini et al. provides a useful methodological foundation, which future studies could extend through joint temporal and geographical testing.

7.3 Early and Progressive Yield Forecasting

Future models should generate predictions at clearly defined stages of the growing season using only information available at each forecast date. Progressive forecasting could update yield estimates as new weather observations, remote-sensing measurements, and crop development records become available. Evaluation should examine the trade-off between forecast lead time and prediction error. Partial-season forecasting experiments in the reviewed literature demonstrate the relevance of assessing this trade-off explicitly.

7.4 Integration of Agricultural Knowledge

A promising direction is to combine ensemble learning with agricultural knowledge concerning crop development, water availability, nutrient requirements, and environmental stress. Candidate approaches include agronomically informed feature construction, constraints on model responses, and integration with process-based crop simulations. These approaches should be compared with purely data-driven baselines to establish whether they improve generalization or reduce implausible predictions. Particular attention should be given to performance when training observations are limited.

7.5 Uncertainty-Aware Prediction

Future studies should complement point estimates with prediction intervals or other clearly defined uncertainty measures. Evaluation should examine interval coverage and width across crops, regions, yield levels, and extreme seasons. Methods should also identify observations that differ substantially from the training distribution. This would help distinguish forecasts supported by familiar conditions from estimates that require greater caution or additional data.

7.6 Stable and Contextual Explanations

Explainability research should move beyond single feature-importance rankings toward assessments of stability across model refits, seasons, and regions. SHAP, permutation importance, PDP, and ALE could be compared to determine when their interpretations agree and why they differ. Because correlated predictors affect explanation methods differently, future studies should document dependence assumptions and reference data. ALE is particularly relevant when conventional partial dependence would evaluate unrealistic predictor combinations.

7.7 Evaluation of Agricultural Decisions

Future work should distinguish explaining a yield estimate from recommending an intervention. Recommendations involving irrigation, fertilizer, or planting practices should be evaluated using agronomic evidence and suitable experimental or causal designs. User studies involving farmers, extension personnel, and agronomists could examine whether explanations are understandable and useful. Field evaluations could then assess whether access to predictions improves decisions, resource use, or economic outcomes.

7.8 Reproducible Benchmarking

Shared benchmarks should provide documented preprocessing procedures, fixed evaluation partitions, simple baselines, and comparable hyperparameter-tuning budgets. Reporting should include MAE, RMSE, R^2 where appropriate, bias, and performance across relevant subgroups. Code, model configurations, and data versions should be made available where permitted. Such practices would make it easier to determine whether improvements arise from the algorithm, additional predictors, or differences in evaluation.

7.9 Efficient Deployment and Model Updating

Future research should assess models under practical constraints, including incomplete inputs, limited connectivity, computational cost, and delayed observations. Comparisons should consider whether simpler ensembles can provide sufficient performance with lower deployment requirements. Operational systems should also monitor changes in input distributions, prediction errors, and explanation patterns. Updating procedures must be evaluated on later observations to establish whether adaptation improves performance without obscuring failures under changing agricultural conditions.

VIII. CONCLUSION

This review examined ensemble learning methods, explainable AI approaches, validation practices, and research challenges in crop yield prediction. The reviewed studies demonstrate that bagging, Random Forest, boosting, and model-combination strategies can capture complex relationships among climatic, soil, and agricultural management variables. However, their relative performance depends on dataset characteristics, predictor availability, forecast timing, and evaluation design. The available evidence therefore does not establish a universally superior method.

Explainability techniques, including feature importance, SHAP, partial dependence, ALE, and LIME, provide complementary ways to examine model behaviour. Their interpretation requires attention to correlated predictors, reference assumptions, and explanation stability. Feature contributions describe predictions and should not be treated as evidence that changing an agricultural input will cause a corresponding change in yield.

Future progress requires representative datasets, reproducible comparisons, realistic temporal and geographical validation, and explicit uncertainty assessment. Integrating agricultural knowledge and evaluating systems with intended users will help connect predictive modelling to practical needs. Reliable crop yield forecasting ultimately depends on balancing accuracy, interpretability, generalization, and the availability of information at the time decisions are made.

REFERENCES

1. Jeong, J. H., et al. (2016). Random Forests for Global and Regional Crop Yield Predictions. *PLOS ONE*, *11*(6), e0156571. https://doi.org/10.1371/journal.pone.0156571
2. Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020). Forecasting Corn Yield With Machine Learning Ensembles. *Frontiers in Plant Science*, *11*, 1120. https://doi.org/10.3389/fpls.2020.01120
3. Shahhosseini, M., Hu, G., Khaki, S., & Archontoulis, S. V. (2021). Corn Yield Prediction With Ensemble CNN-DNN. *Frontiers in Plant Science*, *12*, 709008. https://doi.org/10.3389/fpls.2021.709008
4. Srivastava, A. K., et al. (2022). Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific Reports*, *12*, 3215. https://doi.org/10.1038/s41598-022-06249-w
5. Gupta, S., et al. (2026). SHAP-enhanced ensemble learning for yield prediction: insights from CY-Bench data on Indian wheat. *IAES International Journal of Artificial Intelligence*, *15*(4), 3411–3420. https://doi.org/10.11591/ijai.v15.i4.pp3411-3420
6. Akhtar, M.N.; Shaikh, A.J.; Khan, A.; Awais, H.; Bakar, E.A.; Othman, A.R. Smart Sensing with Edge Computing in Precision Agriculture for Soil Assessment and Heavy Metal Monitoring: A Review. Agriculture **2021**, 11, 475. https://doi.org/10.3390/agriculture11060475
7. Reddy, B. B. ., Syed Gilani Pasha., Kameswari, M. ., Chinkera, R. ., Fatima, S. ., Bhargava, R. & Shrivastava, A. (2024). Classification Approach for Face Spoof Detection in Artificial Neural Network Based on IoT Concepts. International Journal of Intelligent Systems and Applications in Engineering, 12(13s), 79–91. https://ijisae.org/index.php/IJISAE/article/view/4570
8. Patil, B., Babu, D.S., Pasha, S.G., & Shirodkar, V. (2024). A Review of Agricultural IoT: Challenges, Applications & Future Research Direction. International Journal of Management, Technology, and Social Sciences. DOI:10.47992/ijmts.2581.6012.0362