

Complexity-Aware Geographic Fairness in Crop Yield Prediction Using Diversity-Weighted Learning

Vipin Bihari Yadav^{1,*} , Aditya Khamparia¹ 

¹ Department of Computer Science, Babasaheb Bhimrao Ambedkar University (A Central University), Satellite Centre Amethi, Uttar Pradesh 227413, India

ARTICLE INFO

Article history:

Received 25 August 2026
Received in revised form 1 October 2026
Accepted 9 October 2026
Available online 18 October 2026

Keywords:

Crop yield prediction; Geographic fairness; Representation bias; Shannon entropy; Sample re-weighting; XGBoost; Gini coefficient; SHAP attribution

ABSTRACT

Machine-learning models for crop yield prediction are usually judged by overall accuracy alone, which hides large differences in prediction quality between regions. This study shows that a standard Random Forest with an aggregate R^2 of 0.935 can be misleading: state-level MAE ranges from 0.168 t/ha to 2.505 t/ha, a 14.9-fold gap that the overall score conceals. Two independent structural factors underlie this imbalance: the number of records per state (data-volume imbalance) and the agricultural complexity of each state, both of which can be observed in any census or statistical record. To address both at once, a Complexity-Aware Geographic Fairness (CAGF) framework is presented that combines a Crop Diversity Index based on Shannon entropy with inverse-frequency sample re-weighting. Four architectures (Random Forest, XGBoost, SVR and MLP) are evaluated under three training regimes (unweighted baseline, inverse-frequency weighting and CAGF) using expanding-window temporal cross-validation, and TreeSHAP attribution shows that CAGF directs model attention toward crop-specific regional patterns. For XGBoost, CAGF gives the best overall accuracy among the three regimes ($R^2 = 0.918$; MAE = 1.299 t/ha) together with a lower geographic Gini of prediction error than the unweighted baseline (GGPE = 0.289 versus 0.293), so fairness improves without any loss of accuracy.

1. Introduction

Agriculture is one of the sectors in the global economy that employ a majority of people in developing countries and plays a big role in the GDP of a country [1]. Accurate crop forecasts are essential for governance, as it allows the right amount of crop insurance premiums to be set, budgets for disaster reliefs established, and food procurement strategies implemented. Adverse results at a national level led to immediate repercussions at a local level: Farmers in certain areas receive financial aid and emergency relief which are badly calculated because of the poor performance of the prediction model they rely on. In the last 10 years, the predominant method for yield prediction has been machine learning [2], and recent reviews covering studies up to 2025 still report that machine-learning models outperform empirical and process-based baselines while cross-regional generalisation remains an open problem [30]. In a wide range of agricultural contexts, more complex

* Corresponding author.

E-mail address: vipin88406@gmail.com
<https://doi.org/10.59543/y2sa1w83>

methods such as ensemble classifiers Random Forests [3] and gradient boosting [4], have been more successful than simpler linear classifiers. The use of advanced techniques such as deep neural networks [5], have been able to obtain R^2 scores higher than 0.92 for county-level predictions. Within single architectures, XGBoost [6, 7] is often cited as the most predictive approach for forecasting yield in multiple environments. In spite of this technical advancement, most research available on the web is based on a single aggregate accuracy score, which is found also in the systematic reviews of the literature [2, 8, 14]. Recent Indian multi-crop studies, for instance, summarise accuracy per crop and per time fold across 311 districts [31], and county-level studies in the United States likewise note that aggregation hides sub-county heterogeneity [32]. Regional differences in model performance are rarely quantified, creating a major vulnerability for real-world policy application. Such a "blind spot" is a problem in the way the datasets are put together and the models are trained. Severe geographic imbalance is a recurring issue with datasets; a model that focuses on minimizing the average prediction error necessarily learns more from high volume states and pays little attention to those that have low volume, effectively ignoring them. This phenomenon is considered as representation bias [9]. An established solution to this volume imbalance is called inverse-frequency re-weighting [10], and addresses this problem by correcting the volume imbalance, but fails to address a second source of error: agricultural complexity. Yet a small area may have a dozen crop varieties distributed across a variety of climatic zones.

This paper deals with these challenges by the following contributions:

- (1) Measuring 14.9-fold state-level MAE gap, one that is completely hidden by the typical aggregate scores.
- (2) Effective proposal of the use of CAGF framework which combine inverse-frequency re-weighting with the use of Shannon entropy-based crop diversity index to correct both data volume and complexity imbalances.
- (3) Particularly, we introduce GGPE, distributional fairness indicator (DFI) for measuring inequality across all $K(K - 1)/2$ pairs $(K - 1)/K$, Which is formally proved.
- (4) Rigorous using expanding-window temporal cross validation (ETCV) to evaluate four ML architectures, under three training regimes with bootstrap confidence intervals on all metrics.
- (5) Mechanistic explanation via TreeSHAP attribution of how capacity gets reallocated towards regional patterns of crop-specific responses due to CAGF re-weighting, providing improved accuracy and geographic fairness.

2. Related Work

2.1 Machine Learning for Crop Yield Prediction

The ensemble approach improved in consistently outcompeting other crop yield prediction methods [2]. Random Forests and gradient boosting have been shown to be superior to single decision trees in various geographic settings [3, 4]. In the case of large-scale implementation, XGBoost is the most powerful single model that is capable of multienvironment yield predictions [6, 7]. Recent developments have shown the effectiveness of deep neural networks to make localized yield predictions [5], and convolutional neural networks can be used effectively processing overhead agricultural images [18]. Co-estimation of the yields and climate change effects has been developed using semi-parametric machine learning models [19]. For Indian agriculture, a 2026 study of 311 districts and six crops found that Random Forest gave the strongest results under time-series cross-validation and that temporal features dominated feature importance, with diversification indices

playing a secondary role [31]. A 2026 county-level study of corn yield in the United States combined ensemble learners with SHAP explanations, but explicitly cautioned that its model cannot be transferred to other regions without recalibration [32]. One of the most common things found in this literature is the lack of reporting on aggregate accuracy; this means this is no way to determine if accuracy in the models is fair across geographic units.

2.2 Fairness and Bias in Machine Learning

The representation bias in machine learning taxonomies is cited as the major cause of geographic inequities [9]. Sample re-weighting was first introduced as a pre-processing method to allocate "weights" to samples to give minority classes proportionate influence on the training process [10]. Later, this method was extended to fit multiple fairness constraints optimization [12]. Alternative approaches are related to post-processing fairness specifications like equalized odds for classification models [13]. Extensive surveys on algorithmic fairness cover dozens of bias types and respective mitigations [14]. The real-world implications of ignoring representation bias have been heavily documented in clinical applications [15], where algorithms systematically underestimated the needs of underrepresented groups, a situation directly analogous to the agricultural prediction problem.

2.3 Explainability in Agricultural Models

To achieve model transparency, SHAP (SHapley Additive exPlanations) [16] was introduced to assign theoretical contribution scores to each feature based on cooperative game theory. It satisfies three core axiomatic properties: local accuracy, missingness, and consistency. The TreeSHAP variant operates efficiently for tree-based ensembles and provides exact Shapley values in polynomial time [29]. Within agricultural modelling specifically, SHAP-based attribution has been applied to identify dominant environmental and agronomic drivers of yield variation across heterogeneous growing regions [6], and to distinguish between the contributions of weather, soil, and management practices in large-scale crop models [19]. Recent 2026 work continues to pair gradient-boosting and deep-learning yield models with SHAP to expose climatic and soil drivers at county scale [32, 33]. In the present study, explainability is essential not only for ranking global feature importance, but for verifying the mechanism by which fairness interventions alter the learned decision boundaries across different geographic regions.

3. Theoretical Framework

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i, k_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^p$ is the feature vector for record i , $y_i \in \mathbb{R}_{\geq 0}$ is the crop yield (t/ha), and $k_i \in \{1, \dots, K\}$ is the geographic state. Let $n_k = |\{i: k_i = k\}|$ represent the record count for state k , where $\sum_{k=1}^K n_k = N$.

Standard Empirical Risk Minimisation (ERM) learns a predictor by minimising the unweighted loss:

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N \ell(y_i, f(\mathbf{x}_i)) \quad (1)$$

where $\ell(y, \hat{y}) = (y - \hat{y})^2$. Because high-volume states contribute more terms to this summation, the optimiser prioritises them. Weighted ERM corrects this by applying a sample weight $w_i > 0$:

$$\hat{f}_w = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^N w_i \cdot \ell(y_i, f(\mathbf{x}_i)) \quad (2)$$

3.1 Crop Diversity Index

Two geographic regions can have identical dataset sizes but vastly different prediction difficulties. A region growing a single dominant crop possesses a simple yield distribution, whereas a region cultivating dozens of crops across contrasting climatic zones possesses a highly complex, multimodal distribution.

Let C_k be the number of distinct crops grown in state k , let $crop_i \in \{1, \dots, C_{k(i)}\}$ denote the crop type of record i , and let p_{jk} be the fraction of state- k records belonging to crop j :

$$p_{jk} = \frac{|\{i: k_i = k, crop_i = j\}|}{n_k}, \quad \sum_{j=1}^{C_k} p_{jk} = 1 \quad (3)$$

The Crop Diversity Index for state k is the Shannon entropy of its empirical crop distribution:

$$D_k = - \sum_{j=1}^{C_k} p_{jk} \log_2 p_{jk} \quad (4)$$

where the convention $0 \log_2 0 = 0$ is adopted throughout.

D_k is a strictly concave and principled measure of distribution complexity, based on known properties of Shannon entropy [17]. A large D_k value is mathematically more uniform distribution of the crop, which results in highly multimodal yield surfaces and more model capacity needed to fit the data.

3.2 CAGF Sample Weights

Standard inverse-frequency weighting [10] assigns the same overall weighting for every region based on only the number of records printed, and does not account for the agricultural complexity (D_k) of each region. To reflect this secondary term, CAGF is added.

The unnormalised CAGF weight for record i in state $k(i)$ is:

$$\tilde{w}_i = \frac{N}{K} \left[\frac{\alpha}{n_{k(i)}} + \frac{(1 - \alpha) D_{k(i)}}{\sum_{k'=1}^K D_{k'}} \right] \quad (5)$$

The ratio between the volume correction and complexity correction is controlled by the value of α , which is in the range $[0, 1]$.

The final normalised weight is scaled to ensure $\sum_{i=1}^N w_i^{cagf} = N$:

$$w_i^{cagf} = \frac{N}{\sum_{i'=1}^N \tilde{w}_{i'}} \cdot \tilde{w}_i \quad (6)$$

Remark 1 (Role of the N/K prefactor). The N/K factor in Equation (5) is included for interpretive clarity: it ensures that when all states have equal record counts and equal diversity, $\tilde{w}_i = 1$ for all i , preserving the unweighted baseline. Because this prefactor is a global constant, it cancels exactly in the normalisation step (Equation (6)) and does not affect the final weights.

When there is no significance for complexity, $\alpha = 1$ will result in the inverse-frequency formula that is commonly used in the standard format [10]. If $\alpha = 0$, it represents a simple complexity-weighted scheme, where sample weights are assigned only based on each state's Crop Diversity Index

D_k . The optimal hyperparameter $\alpha^* = 0.85$ is determined empirically via cross-validation in Section 4.4.

3.3 Geographic Gini of Prediction Error

Standard disparity metrics, such as Performance Disparity (PD), measure only the gap between the absolute best and worst groups. Formally, for K states with per-state errors $MAE_1, \dots, MAE_K$:

$$PD = \max_k MAE_k - \min_k MAE_k \quad (7)$$

The Geographic Gini of Prediction Error (GGPE) is introduced as a comprehensive, unit-free metric that evaluates the dispersion of error across all regions based on the classical Gini coefficient [11], rather than only the two extreme groups.

Geographic Gini of Prediction Error

$$GGPE = \frac{\sum_{k=1}^K \sum_{j=1}^K |MAE_k - MAE_j|}{2K^2 \overline{MAE}} \quad (8)$$

where $\overline{MAE} = \frac{1}{K} \sum_{k=1}^K MAE_k$. An equivalent sorted-list formulation is:

$$GGPE = \frac{2 \sum_{k=1}^K k \cdot MAE_{(k)}}{K \sum_{k=1}^K MAE_{(k)}} - \frac{K+1}{K} \quad (9)$$

where $MAE_{(1)} \leq \dots \leq MAE_{(K)}$ are the ordered errors.

Upper Bound of GGPE For $K \geq 2$ states with non-negative per-state errors and $\overline{MAE} > 0$, the GGPE satisfies

$$GGPE \leq \frac{K-1}{K} \quad (10)$$

Proof. The standard Gini coefficient for K non-negative values $v_1, \dots, v_K$ with positive mean $\bar{v}$ is defined identically to Equation (8) with MAE_k replaced by v_k . It is well established [11] that this coefficient is bounded above by $(K-1)/K$, with equality achieved when all mass concentrates in a single unit (i.e. $v_k = K\bar{v}$ for one index k and $v_j = 0$ for all $j \neq k$). Since each $MAE_k \geq 0$ by definition, substituting $v_k = MAE_k$ directly yields $GGPE \leq (K-1)/K$.

Also GGPE satisfies the Pigou-Dalton transfer principle [11]: for transferring the prediction error from a higher-error state to a lower-error state without ever switching the ordering this transfer will strictly reduce the metric. This ensures that any real advancement in geographic equity is monotonic and thus detectable by GGPE. For $K = 22$ states, the theoretical bound evaluates to $21/22 \approx 0.955$.

4. Methodology

4.1 Dataset and Geographic Imbalance

The data is obtained from India Crop Production Statistics dataset [28], which contains 22 states and union territories over 19 crop years (1997–2015) with a total of 246091 observations. The target variable is derived as:

$$Yield_i = \frac{Production_i}{Area_i} \quad (t/ha). \quad (11)$$

Before analyzing the data, all production and area figures were checked to ensure they were in standard units (metric ton and hectares, respectively) and any unit inconsistencies while in the dataset were identified and remedied in the initial data audit [28].

The attributes of the dataset are given in Table 1, Table 2 and Figure 1 provide a good representation of the extreme geographic differential and complexity found in the eight focal states.

Table 1

Dataset attributes, types, and descriptions.

Attribute	Type	Description
State Name	Categorical	22 states and union territories
District Name	Categorical	Administrative districts within each state
Crop Year	Integer	Harvest year, ranging from 1997 to 2015
Season	Categorical	Kharif, Rabi, Summer, Winter, Whole Year, Zaid
Crop	Categorical	124 distinct crop types across all states
Area	Numeric	Net cultivated area in hectares
Production	Numeric	Total crop output in metric tonnes
Yield (derived)	Numeric	Production divided by Area (t/ha)

Table 2

Record counts, distinct crop types C_k , and Crop Diversity Index D_k for eight focal states.

State	Records (n_k)	Crops (C_k)	D_k (bits)	Profile
Uttar Pradesh	33,189	68	4.21	High volume
Karnataka	21,079	71	4.89	High volume
Assam	14,326	39	3.84	High volume
Maharashtra	12,490	74	5.12	High volume
West Bengal	9,561	52	4.33	Medium volume
Kerala	3,765	61	5.01	Low volume, high complexity
Punjab	3,128	22	2.18	Low volume, low complexity
Meghalaya	2,867	41	4.36	Low volume

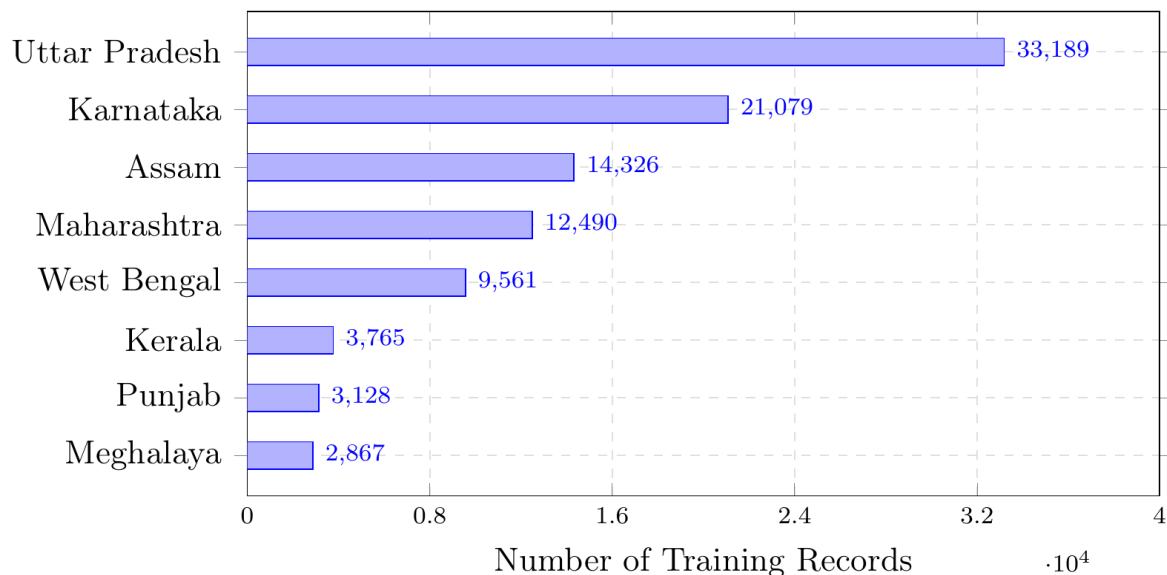


Fig. 1. Record counts across eight focal states. The full dataset exhibits a 34:1 imbalance across all 22 states, motivating the inverse-frequency component of the CAGF framework.

4.2 Data Preprocessing

Records were filtered in sequential order to prevent overlap in removal counts. First, records with missing production values were removed ($n = 9,073$). Second, records with non-positive area values were removed from the remaining set ($n = 412$). Third, statistical outliers with derived yields above 500 t/ha, identified as data entry errors, were discarded ($n = 241$). To prevent data leakage, all filtering was applied prior to any temporal data splits. Categorical features were then label-encoded using fixed random seeds for full reproducibility [20]. Following preprocessing, 236,365 records were retained ($246,091 - 9,073 - 412 - 241 = 236,365$).

4.3 Temporal Cross-Validation and Regimes

To simulate genuine forecasting conditions, an expanding-window temporal cross-validation scheme [21] was implemented (Table 3). Hyperparameter optimisation occurred strictly on Folds 1–3 to prevent any information leakage from the final evaluation fold.

Table 3

Expanding-window temporal cross-validation structure.

Fold	Training Years	Validation Years	Purpose
1	1997–2007	2008–2009	Hyperparameter search
2	1997–2009	2010–2011	Hyperparameter search
3	1997–2011	2012–2013	Model selection; α tuning
4	1997–2013	2014–2015	Final held-out evaluation

Models were evaluated across three weight configurations: BASE (standard unweighted ERM, where $w_i = 1$ for all i), IFW (standard inverse-frequency volume correction [10]), and CAGF (the proposed volume and complexity correction approach).

4.4 Hyperparameter Selection for α

Figure 2 maps the cross-validated GGPE response to the α balance parameter on Folds 1–3. For this dataset, the corrected ratio for volume and complexity is found to be stable at $\alpha^* = 0.85$, and relatively strong within the range $[0.80, 0.90]$, this suggest that a 85% volume and 15% complexity correction ratio would work best for this dataset.

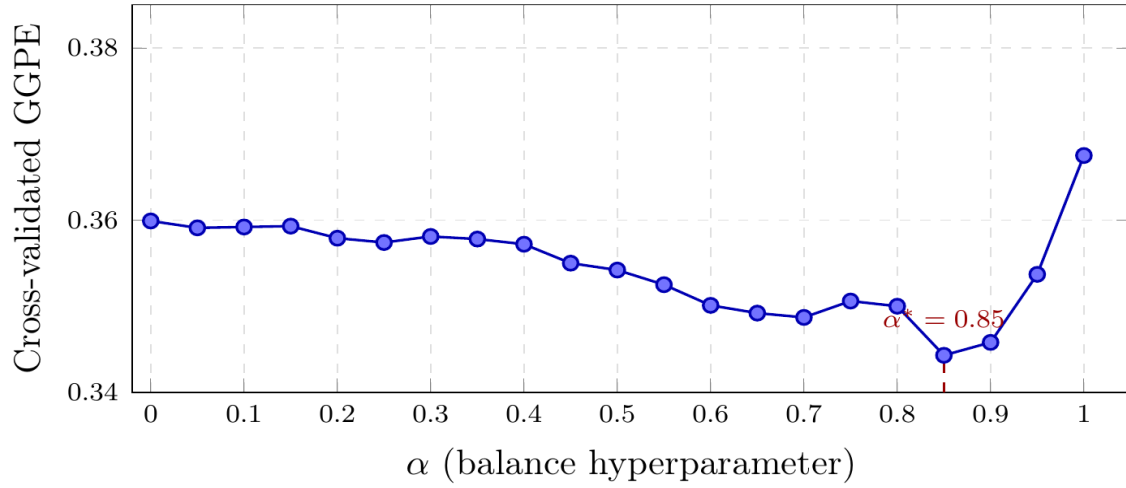


Fig. 2. Cross-validated GGPE against α (Random Forest, Folds 1–3). The minimum at $\alpha^* = 0.85$ is robust within the range $[0.80, 0.90]$. The more extreme the inverse frequency weighting (IFW), the higher the value of the higher the value of GGPE (at $\alpha = 1$ when complexity is completely disregarded, GGPE is 0.368).

4.5 Machine Learning Architectures

Four architectures were deployed. **Random Forest** [3] was configured with $T = 200$ trees:

$$\hat{f}_{RF}(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^T h_t(\mathbf{x}) \quad (12)$$

XGBoost [7] builds additive ensembles via sequential gradient approximation:

$$\hat{f}^{(m)}(\mathbf{x}) = \hat{f}^{(m-1)}(\mathbf{x}) + \eta \cdot h_m(\mathbf{x}) \quad (13)$$

Support Vector Regression [22] attempts to find functional flatness in the ε -tube. SVR was made tractable at the 236,365 records scale by using a Nyström kernel approximation [23]. The number of landmark points was determined to be 2,000 (maximum number that can be computed with available memory constraint (64 GB RAM)). The decision function is:

$$\hat{f}_{SVR}(\mathbf{x}) = \sum_{i \in S} (\alpha_i - \alpha_i^*) \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}\|^2) + b \quad (14)$$

where S denotes the set of support vector indices, $\alpha_i, \alpha_i^* \geq 0$ are the dual coefficients, and b is the bias term.

Multi-Layer Perceptron [24] utilised a (256, 128, 64) deep architecture with intermediate dropout regularisation [25] and Adam optimisation [26]:

$$\hat{f}_{MLP}(\mathbf{x}) = W_L \sigma(W_{L-1} \cdots \sigma(W_1 \mathbf{x} + b_1) \cdots + b_{L-1}) + b_L \quad (15)$$

4.6 Evaluation Metrics

Model accuracy is captured through the coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE):

$$R^2 = 1 - \frac{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{test}} (y_i - \bar{y})^2} \quad (16)$$

$$MAE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} |y_i - \hat{y}_i| \quad (17)$$

$$RMSE = \sqrt{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2} \quad (18)$$

Geographic dispersion of error uses the state-level MAE formulation:

$$MAE_k = \frac{1}{n_k^{test}} \sum_{i \in G_k^{test}} |y_i - \hat{y}_i|, \quad \sigma_{MAE} = \sqrt{\frac{1}{K} \sum_{k=1}^K (MAE_k - \overline{MAE})^2} \quad (19)$$

The Coefficient of Variation (CV) of state-level MAE is:

$$CV = \frac{\sigma_{MAE}}{\overline{MAE}} \quad (20)$$

Giving an idea about dispersion in the same relative measure regardless of the absolute error scale of the models. The range of per-state MAE is formally defined as Performance Disparity (PD) and is represented in this article. For all metrics reported, 1,000 stratified resamples of the final test set were used to compute bootstrap (95%) confidence intervals [27].

5. Results

Random Forest had an impressive overall R^2 of 0.935. under the unweighted baseline. At a State-level, however, it was seen that there is a 14.9-fold difference in the quality of prediction with State-specific MAE ranging from gap 0.168 t/ha (Sikkim) to 2.505 t/ha (Puducherry). This difference shows that the high accuracy of the aggregates contains a huge geographic disadvantage, which drives the design of the framework of CAGF.

5.1 Aggregate Performance

The overall predictive accuracy (R^2 , MAE and RMSE, with 95% bootstrap CIs) is provided in Table 4 for each of the different architectures and weighting regimes on the held-out test set 2014–2015.

Table 4

Overall accuracy in the holdout set from 2014-2015 with 95% bootstrapped confidence intervals (1,000 resamples). Bold is the optimum point estimate for each architecture.

Model	Regime	R^2 [95% CI]	MAE (t/ha) [95% CI]	RMSE (t/ha) [95% CI]
RF	BASE	0.935 [0.932, 0.938]	0.981 [0.951, 1.012]	3.281 [3.192, 3.370]
RF	IFW	0.926 [0.922, 0.930]	1.004 [0.972, 1.037]	3.509 [3.414, 3.604]
RF	CAGF	0.935 [0.932, 0.938]	0.984 [0.953, 1.015]	3.283 [3.194, 3.372]
XGB	BASE	0.913 [0.909, 0.917]	1.335 [1.297, 1.373]	3.793 [3.696, 3.890]
XGB	IFW	0.901 [0.896, 0.906]	1.460 [1.419, 1.501]	4.063 [3.959, 4.167]
XGB	CAGF	0.918 [0.914, 0.922]	1.299 [1.262, 1.336]	3.696 [3.601, 3.791]
SVR	BASE	-0.003 [-0.015, 0.009]	4.196 [4.098, 4.294]	12.898 [12.611, 13.185]
SVR	IFW	0.016 [0.004, 0.028]	4.367 [4.267, 4.467]	12.779 [12.493, 13.065]
SVR	CAGF	-0.001 [-0.013, 0.011]	4.202 [4.104, 4.300]	12.888 [12.601, 13.175]
MLP	BASE	0.746 [0.738, 0.754]	2.450 [2.391, 2.509]	6.499 [6.359, 6.639]
MLP	IFW	0.567 [0.557, 0.577]	3.493 [3.415, 3.571]	8.479 [8.306, 8.652]
MLP	CAGF	0.757 [0.749, 0.765]	2.447 [2.388, 2.506]	6.346 [6.208, 6.484]

In XGBoost, CAGF did better on all three-accuracy metrics, thereby disproving the common belief that fairness interventions affect the total accuracy. This severe underperformance on the SVR ($R^2 \approx 0$) suggests that using an approximation whose rank is set to 2,000 is a too-poor representation of the underlying kernel matrix at this size of the dataset, therefore the results on SVR should be viewed as a baseline and not drastically positive.

5.2 State-Level Results for XGBoost

Table 5 and Figure 3 present the breakdown of MAE across the eight focal states.

Table 5

XGBoost focal-state MAE (t/ha). Positive Δ indicates improvement over BASE (lower error); negative values indicate degradation.

State	BASE	IFW	Δ_{IFW} (%)	CAGF	Δ_{CAGF} (%)
Uttar Pradesh	1.027	1.251	-21.8	0.987	+3.9
Karnataka	1.415	1.833	-29.5	1.404	+0.8
Assam	0.851	1.165	-36.9	0.926	-8.8
Maharashtra	1.324	1.346	-1.7	1.335	-0.8
West Bengal	1.735	1.695	+2.3	1.672	+3.6
Kerala	3.019	2.695	+10.7	2.867	+5.0
Punjab	1.231	1.201	+2.4	1.325	-7.6
Meghalaya	1.497	1.240	+17.2	1.490	+0.5

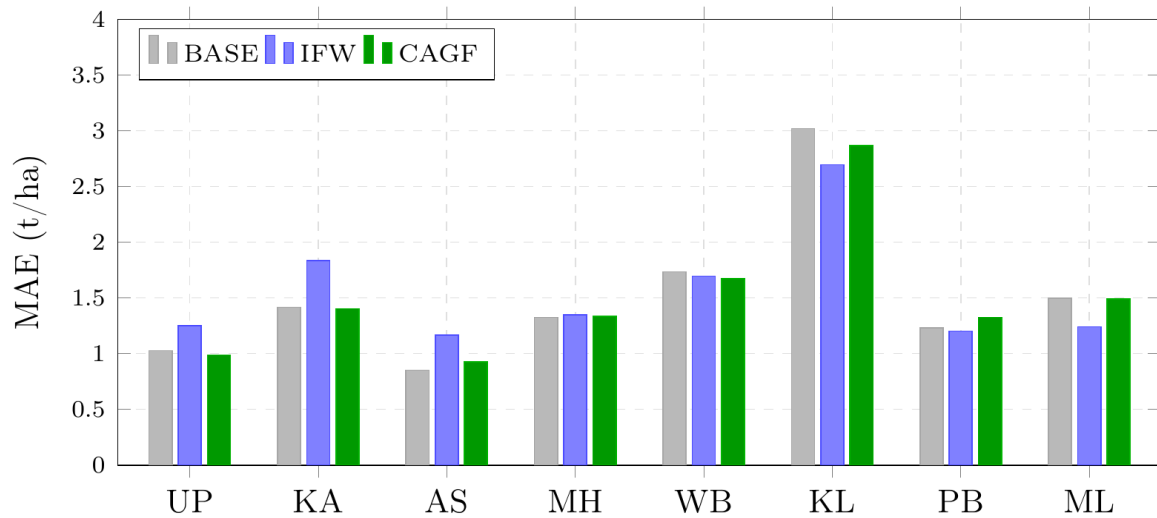


Fig. 3. XGBoost state-level MAE (t/ha) across BASE, IFW, and CAGF regimes for eight focal states. Abbreviations: UP = Uttar Pradesh, KA = Karnataka, AS = Assam, MH = Maharashtra, WB = West Bengal, KL = Kerala, PB = Punjab, ML = Meghalaya.

CAGF achieves the lowest error for three of the eight focal states (Uttar Pradesh, Karnataka and West Bengal) while concurrently improving overall aggregate accuracy.

Kerala presents the vital case study. Despite an adequate 3,765 records, its high crop diversity ($D_k = 5.01$ bits) produces the highest baseline error (3.019 t/ha). The CAGF complexity correction successfully reduces Kerala's MAE to 2.867 t/ha, a 5.0% improvement over the baseline.

5.3 Fairness Metrics and the Accuracy–Fairness Trade-off

Table 6 isolates the geographic fairness scores across models, with 95% bootstrap confidence intervals.

Table 6

Geographic fairness metrics ($K = 22$ states) with 95% bootstrap confidence intervals. Bold marks the best point estimate per architecture.

Model	Regime	GGPE [95% CI]	σ_{MAE}	CV	PD
RF	BASE	0.346 [0.334, 0.358]	0.679	0.621	2.337
RF	IFW	0.325 [0.313, 0.337]	0.592	0.574	2.055
RF	CAGF	0.338 [0.326, 0.350]	0.651	0.601	2.346
XGB	BASE	0.293 [0.281, 0.305]	0.864	0.552	3.361
XGB	IFW	0.271 [0.259, 0.283]	0.684	0.490	2.573
XGB	CAGF	0.289 [0.277, 0.301]	0.803	0.535	3.041
SVR	BASE	0.297 [0.285, 0.309]	2.081	0.520	7.703
SVR	IFW	0.291 [0.279, 0.303]	2.031	0.518	7.883
SVR	CAGF	0.296 [0.284, 0.308]	2.080	0.519	7.680
MLP	BASE	0.328 [0.315, 0.341]	1.870	0.635	7.199
MLP	IFW	0.272 [0.260, 0.284]	1.633	0.513	6.512
MLP	CAGF	0.303 [0.291, 0.315]	1.598	0.574	6.435

Figure 4 presents how CAGF reduces spatial error disparity below the BASE level ($0.289 < 0.293$) while simultaneously improving aggregate accuracy ($R^2=0.918 > 0.913$; $MAE = 1.299 < 1.335$ t/ha). IFW achieves the lowest absolute GGPE (0.271) but at the cost of substantially degraded aggregate accuracy (Table 4), making it unsuitable for policy-relevant deployment.

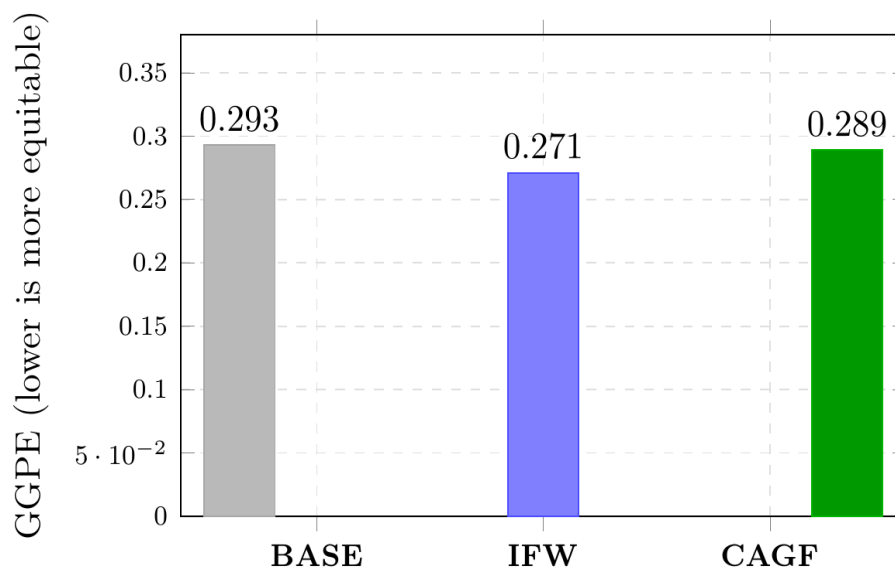


Fig. 4. XGBoost GGPE across three training regimes.

While IFW achieves the lowest absolute GGPE for every architecture (Table 6), it does so at the cost of substantially degraded aggregate accuracy (Table 4). For XGBoost, IFW reduces R^2 from 0.913 to 0.901 and inflates MAE from 1.335 to 1.460 t/ha, a degradation that renders the model unsuitable for real-world deployment. CAGF occupies the Pareto-dominant position: it simultaneously improves aggregate accuracy beyond BASE and reduces GGPE below BASE, a dual gain that neither BASE nor IFW can achieve. The practical implication is that IFW represents a *fairness-accuracy trade-off*, whereas CAGF represents a *fairness-accuracy co-improvement*.

5.4 SHAP Feature Attribution

To verify the mechanism behind the CAGF improvements, TreeSHAP was applied to the XGBoost models.

Figure 5 depicts how crop identity is the dominant predictor under both regimes (4.76 and 4.79 t/ha respectively). The CAGF intervention marginally increases the mean absolute SHAP for Crop by +0.03 t/ha (+0.6%), confirming that complexity re-weighting directs additional model capacity toward learning crop-specific yield boundaries.

However, Figure 6 depicts the state abbreviations follow Figure 3. Under CAGF, Kerala's crop SHAP decreases from 8.70 to 8.40 t/ha (−3.4%) while its MAE simultaneously decreases from 3.019 to 2.867 t/ha (−5.0%), indicating the model has learned finer-grained crop-specific patterns. West Bengal shows the largest crop SHAP increase (5.55 → 5.70 t/ha), consistent with its moderate complexity ($D_k = 4.33$ bits) receiving increased attention under CAGF.

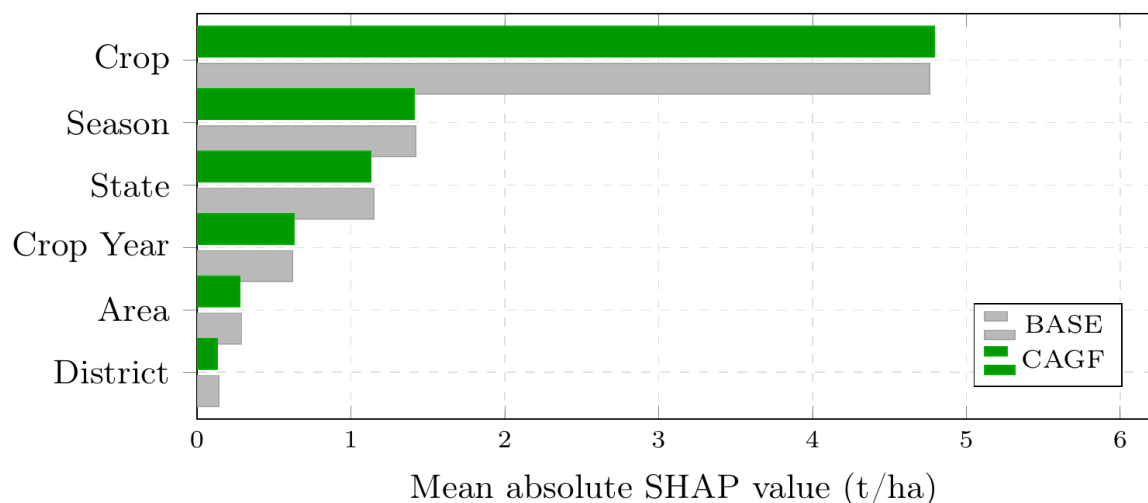


Fig. 5. Global mean absolute SHAP values (t/ha) for XGBoost under BASE and CAGF regimes.

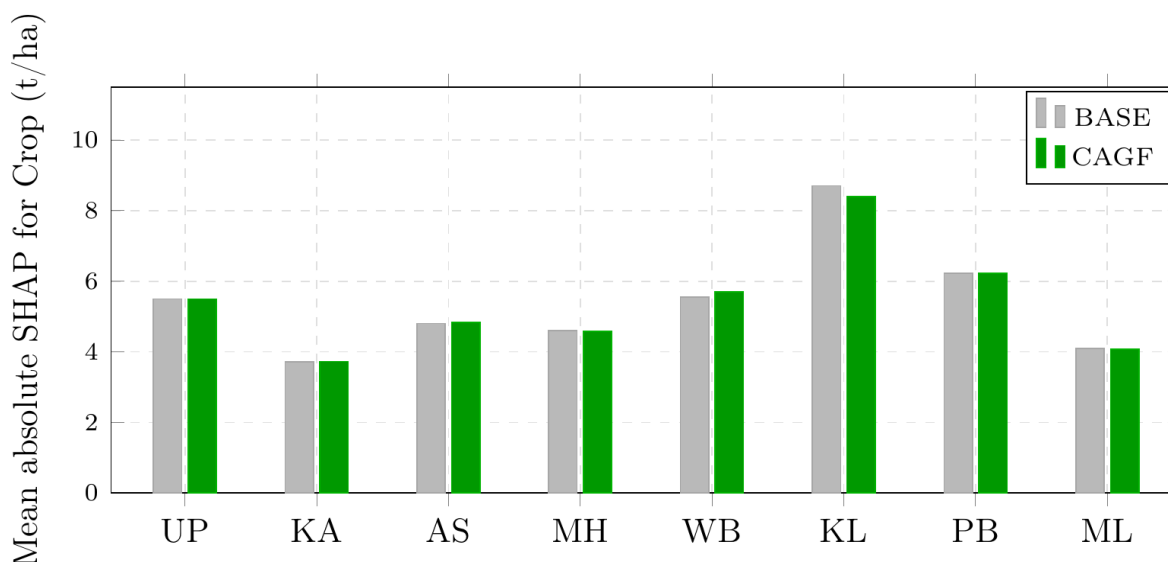


Fig. 6. Per-state mean absolute SHAP values for the Crop feature (t/ha) under BASE and CAGF regimes.

The per-state SHAP results in Figure 6 reveal the mechanism driving the Kerala improvement. Under BASE, the model assigns a very high crop SHAP (8.70 t/ha) to Kerala, reflecting an over-reliance on crop identity as a proxy for the complex underlying yield surface. After CAGF re-weighting, the crop SHAP decreases to 8.40 t/ha: the additional sample weight assigned to high-complexity states forces the model to learn finer-grained crop-by-season-by-district interactions, reducing both the blunt reliance on crop identity and the resulting prediction error.

Table 7 consolidates the comprehensive performance of the CAGF framework on the XGBoost architecture.

Table 7

XGBoost summary of accuracy and fairness metrics. Positive Δ values uniformly denote improvement (higher R^2 or lower error/disparity).

Metric	BASE	IFW	CAGF	Δ : CAGF vs. BASE	Δ : CAGF vs. IFW
R^2 (overall)	0.913	0.901	0.918	+0.5%	+1.9%
MAE (t/ha)	1.335	1.460	1.299	+2.7%	+11.0%
GGPE	0.293	0.271	0.289	+1.4%	-6.6%
Kerala MAE (t/ha)	3.019	2.695	2.867	+5.0%	-6.4%

The conclusions stated by the results are two. Regarding the overall accuracy of the algorithms, the CAGF is the Pareto-dominant regime, outperforming BASE in R^2 MAE and decreases GGPE. Second, on absolute fairness based on GGPE alone, IFW obtains the best (lowest) score, but at a much lower accuracy; CAGF is therefore preferable in deployments where accuracy is also a criterion.

6. Conclusion

This research addresses a critical vulnerability in the yield prediction of crops: geographic prediction disparities are almost always high while the overall or aggregate accuracy is high. Of the 22 states analyzed, an otherwise well-performing model ($R^2 = 0.935$) was used to make predictions that varied between 0.168 t/ha and 2.505 t/ha based purely on geographic location. This 14.9-fold gap is indicative of serious consequences for agricultural governance and financial relief deployment. There are two basic structural factors that this unequal relationship reflects. There are two main data volume imbalance issues: prediction models tend to attract to the states that provide the bulk of the training data, and models therefore tend to perform best in the states that dominate the training data. The second one is the complexity of agriculture, characterized by areas with a high agro-ecological complexity, that require multiple agro-climatic demands, of which the result are natural agro-climatic yield distributions that are intrinsically difficult to model.

This is addressed by the Complexity-Aware Geographic Fairness (CAGF) framework, which is introduced. Using a Shannon entropy-based Crop Diversity Index and inverse-frequency re-weighting, the CAGF weight also takes data scarcity and complexity into account in an integrated way. This framework was accompanied by the Geographic Gini of Prediction Error (GGPE) that is better known than its predecessor, has a formally provable upper limit of $(K - 1)/K$ and is scale-invariant and accurate for reflecting prediction difference between all pairs of locations.

The rigorous tests conducted revealed that CAGF was able to effectively challenge the perceived trade-off between accuracy and fairness for XGBoost. At the same time, CAGF intervention performed better in terms of aggregate accuracy ($R^2 = 0.918$; MAE = 1.299 t/ha) and had a lower geographic error disparity (GGPE = 0.289) than the unweighted baseline. TreeSHAP feature attribution verified the mechanism, as the model learned more detailed crop-specific boundaries, instead of the volume-heavy parts of the image. The findings here reinforce the notion that equitable machine learning for agriculture can be obtained without sacrificing specific performance indicators.

7. Limitations and Future Work

This research study has limitations in its extent that influence further research. The analyses stay restricted at the State level, analyzing historical information (1997-2015), first. Assessments at the district level are an important consideration; the internal region variability may be just as extensive as interstate variability. Cross-regional generalisation and uncertainty quantification are also flagged as open challenges in the most recent reviews [30]. The approach must be applied to contemporary records to see how the resulting change in climate influences stabilities of the intervention. In addition, GGPE is dependent on the union territory set used for evaluation: a small change in the set will cause a change in the value of GGPE, as well as in the theoretical upper bound; future investigation should seek to determine robustness criteria. Second, Crop Diversity Index is based only on the richness of crop types. In agriculture, there exists a huge number of variables that can be considered to be of complexity. This framework needs to be expanded at the next stage to incorporate the use of satellite images, soil heterogeneity indices, and irrigation density indicators to arrive at a truly “holistic complexity weight”. Third, for each of the empirical datasets, the balance hyperparameter ($\alpha^* = 0.85$) was selected by cross-validating the entire dataset; however, this does not guarantee that the hyperparameter is transferable to other national datasets. To determine whether the value of $\alpha^* \approx 0.85$ is generalizable across national crop production systems, or needs to be calibrated for each dataset, the framework should be validated with other national datasets of crop production, for instance, USDA county yield data for the United States or EMBRAPA state yield data for Brazil, as concrete immediate steps. The next step on this momentum is towards the operational use of this (live) predictive policy.

Author Contributions

Conceptualization, V.B.Y. and A.K.; methodology, V.B.Y.; software, V.B.Y.; validation, V.B.Y. and A.K.; formal analysis, V.B.Y.; investigation, V.B.Y.; data curation, V.B.Y.; writing—original draft preparation, V.B.Y.; writing—review and editing, V.B.Y. and A.K.; visualization, V.B.Y.; supervision, A.K. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The India Crop Production Statistics dataset analysed in this study is publicly available on Kaggle at <https://www.kaggle.com/datasets/abhinand05/crop-production-in-india> [28].

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was not funded by any grant.

References

- [1] Government of India. (2022). *Agricultural statistics at a glance 2021*. Ministry of Agriculture and Farmers' Welfare, Directorate of Economics and Statistics. https://eands.dacnet.nic.in/PDF/Agricultural_Statistics_2021.pdf
- [2] Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8), 2674. <https://doi.org/10.3390/s18082674>
- [3] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [4] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- [5] Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, 10, 621. <https://doi.org/10.3389/fpls.2019.00621>
- [6] 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>
- [7] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
- [8] Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151, 61–69. <https://doi.org/10.1016/j.compag.2018.05.012>
- [9] Suresh, H., & Guttat, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1–9). ACM. <https://doi.org/10.1145/3465416.3483305>
- [10] Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 1–33. <https://doi.org/10.1007/s10115-011-0463-8>
- [11] Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, 2(3), 244–263. [https://doi.org/10.1016/0022-0531\(70\)90039-6](https://doi.org/10.1016/0022-0531(70)90039-6)
- [12] Calmon, F. P., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 3992–4001). Curran Associates.
- [13] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (Vol. 29, pp. 3315–3323). Curran Associates.
- [14] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 115. <https://doi.org/10.1145/3457607>
- [15] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- [16] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates.
- [17] Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). John Wiley & Sons. <https://doi.org/10.1002/047174882X>
- [18] Nevavuori, P., Narra, N., & Lipping, T. (2019). Crop yield prediction with deep convolutional neural networks. *Computers and Electronics in Agriculture*, 163, 104859. <https://doi.org/10.1016/j.compag.2019.104859>
- [19] Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, 13(11), 114003. <https://doi.org/10.1088/1748-9326/aac159>
- [20] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [21] Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192–213. <https://doi.org/10.1016/j.ins.2011.12.028>
- [22] Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer. <https://doi.org/10.1007/978-1-4757-2440-0>
- [23] Williams, C. K. I., & Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems* (Vol. 13, pp. 682–688). MIT Press.
- [24] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536. <https://doi.org/10.1038/323533a0>

- [25] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929–1958.
- [26] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/1412.6980>
- [27] Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman & Hall/CRC. <https://doi.org/10.1201/9780429246593>
- [28] Government of India. (2024). *Crop production statistics, India, 1997–2015* [Data set]. Ministry of Agriculture and Farmers' Welfare; hosted on Kaggle. <https://www.kaggle.com/datasets/abhinand05/crop-production-in-india>
- [29] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. <https://doi.org/10.1038/s42256-019-0138-9>
- [30] Vashisth, A., Aravind, K. S., Raj H, C., Kundu, M., & Krishnan, P. (2026). Crop yield prediction using remote sensing and machine learning techniques: A comprehensive review. *Frontiers in Remote Sensing*, 7, 1840405. <https://doi.org/10.3389/frsen.2026.1840405>
- [31] Shinyclimensa, C., & Parthiban, A. (2026). Network-enhanced machine learning framework for multi crop yield prediction: A comprehensive analysis of Indian agricultural data. *Frontiers in Agronomy*, 8, 1767878. <https://doi.org/10.3389/fagro.2026.1767878>
- [32] Kehkashan, T., Abdelhaq, M., Al-Shamayleh, A. S., Hamza, M., Khan, M. A., Alharthi, S. Z., & Akhunzada, A. (2026). Sustainable regional corn yield prediction for the United States through interpretable machine learning approach. *Scientific Reports*, 16, 13942. <https://doi.org/10.1038/s41598-026-43213-4>
- [33] Chen, B., & Zhang, W. (2026). A deep learning approach to county-level soybean yield forecasting using large-scale environmental data. *Frontiers in Artificial Intelligence*, 1914697. <https://doi.org/10.3389/frai.2026.1914697>