

A Reliability-Aware Distributionally Invariant and Uncertainty-Selective Framework for Multi-Regional Crop Yield Classification

Phanikanth Chintamaneni¹, G. Krishna Mohan², Kodukula Subrahmanyam³

¹Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.

²Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.

³Professor, Department of Computer Science and Engineering, Dr.RVR NRI Institute of Technology Deemed to be University, Pothavarappadu (M), Agiripalli (M), Andhra Pradesh, India.

Abstract

Agricultural Internet-of-Things data differ greatly between farms, regions, seasons, and sensor deployments. A predictor learned on pooled data can have high average accuracy but break when missingness, sensor drift, class imbalance or geographic covariate shift modifies the input distribution. In this paper, we introduce RADIUS-Yield, a novel extension for low, medium, and high crop-yield classification. The framework substitutes discrete hard filtering with continuous reliability weights, substitutes entropy-only sampling with a class-region constrained coresets, connects regional experts via a similarity graph, modifies global and local probabilities via a shift gate, and yields conformal prediction sets with an optional rejection option. Five algorithms follow the architecture sequence, and five propositions prove bounded reliability weights, monotone class-region coverage, convex graph regularization for fixed representations, probability simplex preservation, and marginal conformal coverage under exchangeability. Raw field records were not accessible for a new run. A seeded controlled experiment was simulated with 18,000 instances, 24 features, six regions, 1,800 farm groups, region-dependent missingness, sensor faults, nonlinear class structure, and an induced test shift. RADIUS-Yield achieved a macro F1 of 0.7500 and an AUROC of 0.9174 with an expected calibration error of 0.0219 and a farm-bootstrap 95% macro-F1 interval of [0.7361,0.7631]. Multinomial logistic regression resulted in the best macro F1 at 0.7583, so the experiment does not support the assertion of a method being superior in general. Under the enforced shift, the 90% conformal set attained 88.63% coverage, revealing the limit of exchangeability-based calibration. The results imply the internal functioning of the novel reliability, regional, and selective mechanisms and delineate the field-data experiments required prior to making deployment assertions.

Index Terms—Calibration, conformal prediction, crop-yield classification, distribution shift, Internet of Things, regional experts, selective prediction.

1. Introduction

Crop-yield classification assists in irrigation management, input distribution, storage planning, insurance underwriting and aggregate harvest estimation. Agricultural sensing networks include soil moisture, soil chemistry, temperature, rainfall, vegetation indices, irrigation logs, crop type, and management practices. Such measurements do not result from a single stationary process. Sensor brands are different, maintenance policies are not the same, the farms have different practices, and more in particular regional climate regimes transform the relation of a measurement to a final yield class. A model can be well-fit to a pooled development set, and yet it can perform poorly when it is applied to data from a new farm, a different season, or a new device. Regional heterogeneity needs to be modeled as part of the prediction task rather than as a nuisance to be eliminated through global normalization [1]–[5].

Previous crop yield prediction systems employ random forests, support vector machines, multilayer perceptron, gradient boosting, and ensemble combinations [6]–[17]. These approaches are capable of learning nonlinear interactions and typically outperform linear baselines when development and test observations are drawn from similar distributions. Their reported results apply to different crops, spatial units, definitions of the targets, and splitting rules, so the reported figures cannot be aggregated into a single consolidated ranking. The literature still provides some useful design insights: nonlinear relationships are important, regional context is critical, various learner families may be complementary, and predictions for planning should be accompanied by uncertainty.

The previous HPRF-EWEL model used weak and strong filters at the first two levels of filtering prior to entropy balancing and regional ensemble learning. This design improved the organization of noisy instances, but hard thresholds induce a discontinuity: two almost equal samples nearby a threshold are treated very differently. An accepted sample receives full influence and a rejected sample receives none. That same rule might eliminate an

unusual but real soil, climatic, and management scenario. Entropy accounts for the class histogram, but it does not account for regional coverage, covariate shift, sensor reliability, or error calibration. A fixed regional ensemble may lose effectiveness if the operating point is distant from the local training distribution in the feature space.

RADIUS-Yield changes the central hypothesis. The question is no longer whether a cascade can eliminate uncertain instances prior to the classification. The new hypothesis is that continuous sample reliability, constrained class-region coverage, graph-connected regional experts, and calibrated selective decisions may result in more stable behaviour under a controlled shift than a hard-filtered ensemble. The model keeps all finite observations, apportions influence via a bounded quality weight, protects rare class-region cells through explicit constraints, and modulates the global-local mixture via the distance between a query and its regional training centroid.

The study is structured around five research questions. RQ1 is designed to investigate whether continuous reliability weighting can suppress the effect of missing or outlying sensor records without excluding rare but otherwise valid cases. RQ2 is designed to investigate whether a distribution-constrained coreset maintains every observed class-region cell with a fixed training budget. RQ3 is designed to investigate whether graph regularization can facilitate information transfer between related regions without forcing identical predictors. RQ4 is designed to investigate whether a shift gate and temperature scaling can improve calibration and maintain competitive discrimination. RQ5 is designed to investigate whether conformal prediction sets and selective decisions reveal uncertainty under regional shift, instead of hiding it behind a single class label.

This is a mechanistic contribution. The reliability score is a composite of missingness, median absolute deviation, and regional shift, and is transformed into a continuous weight. The coreset objective combines class-region lower bounds with uncertainty and density terms. A regional graph links locations via standardized centroid similarity and regularizes expert parameters using a Laplacian penalty. A shift gate mixes each regional expert with a global expert based on query distance. Calibration partition learns one temperature and one conformal quantile without reusing the test labels. For each mechanism, there is a corresponding removal experiment and a diagnostic.

This study is based on simulated data rather than real-world farm observations. The data generator matches the source study's dimensions, but not the original dataset's unknown distribution. Farms are kept intact when dividing the data into training and test sets, and the test set is deliberately given a different mix of input values. Sensor errors and region-specific gaps are added before missing values are filled in. These choices let us check whether the equations and methods work as expected. They do not show that the results are agronomically sound, useful in practice, or applicable to the source database of one million records.

The reported ranking is by design evidence-led. RADIUS-Yield did not obtain the best macro F1. Logistic regression was in addition stronger at that endpoint, indicating that the controlled label boundary had a significant additive term. RADIUS-Yield outperforms random forest and is competitive to support-vector in the AUROC, yields small calibration error, and produces small conformal sets. The coverage deficit under the imposed shift is a useful failure result: exchangeability-based calibration needs grouped or shift-aware adjustments before prediction sets can carry a nominal guarantee in a shifted region.

We present the evidence that motivates the problem, the formal problem definition, model, algorithms, relevant mathematical properties, controlled evaluation protocol, results, and limitations in the paper. Architecture terms are consistent in the equations, pseudocode, ablations, figures, and discussion. This allows every claimed contribution to be verifiable and avoids a standard learner being rebranded as a novel approach.

II. Related Work

Crop-yield prediction is a multidisciplinary problem that involves agronomy, environmental sensing, remote observation, and statistical learning. Previous studies at the county and farm level have used a variety of soil, weather, phenology, management, and vegetation characteristics with tree ensembles, support-vector models, neural networks, and stacked predictors [1], [6], [10], [12]. Remote-sensing studies provide improved spatial coverage but are limited by cloud contamination, revisit periods, changes in sensors, and discrepancies between the spectral response and field measurements [4], [5], [8], [14]–[17]. IoT tables provide finer local measurements, but the data have missing packets, calibration drift, duplicated records, and correlated samples from the same farm.

Random forests continue to be a good baseline for nonlinear tabular data and for interactions between variables [18]. Support-vector machines provide a margin-based comparison and can produce calibrated probabilities

following a separate mapping stage [19], [27]. Mixture models capture multimodal class distributions using latent components [20], as extreme learning machines provide a specific randomized hidden-layer analogue [23]. These model families encode different assumptions, motivating their ensemble use [31], [32]. A mixture is meaningful only if the weights are inferred without test leakage and if correlated errors are taken into account.

Class imbalance is typically handled with resampling, synthetic observations, class weights, or threshold changes. SMOTE generates points between minority neighbors [22], but an interpolated point may correspond to an agronomically implausible soil-climate combination. Global entropy is a measure of the uniformity of class counts [21], however equal counts globally may be accompanied by an empty high-yield class in one region. RADIUS-Yield sets lower bounds for every observed class-region cell and never produces a missing cell. The coreset of RADIUS-Yield is an indexed subset of the observed or controlled data, where the provenance of each selected item is preserved.

Methods of quality control in addition tend to discard records that are above a predetermined anomaly threshold. Culling is appropriate for nonfinite or physically unrealistic observations, but model uncertainty is not the same as data invalidity. Since the training set does not contain similar examples, a rare record can be predicted with low confidence. Continuous weighting distinguishes influence from existence: a low-quality record still contributes, but to a much lesser extent, to parameter estimation, but remains accessible for spatial coverage and auditing functions. The score is based on medians and median absolute deviations, derived from the training, making it less sensitive to outliers.

Regional learning can be categorized into two main types: a single pooled model for all locations or separate independent models for each location. Using a pooled model provides more statistical support but may obscure important local relationships. Independent models retain local characteristics but can become unstable when a region has only a few samples. Graph regularization offers a compromise [38]. In this case, each region is a node, similarity between regions is an edge, and a Laplacian penalty is used to prevent large parameter differences between strongly connected nodes. However, the penalty does not force all of the experts to be the same; a weak or non-existent edge allows different regional behaviors.

Distribution shift alters the distributions of input, class prior, conditional label, or sensor representation [37]. Being far from a training centroid does not indicate prediction error, but it does provide a clear signal for modulating the trust in a local expert. The introduced gate is computed utilizing normalized feature distance and a limited mixing. A far query is more influenced by the global learner with a larger support. A close query is more influenced from the regional expert. The gate is saved with each prediction so you can audit regional behavior.

Probability calibration relates to how closely the reported confidence levels correspond to the empirical frequencies. Platt calibration and later calibration research established that there exist post-hoc methods to transform scores into probabilities [27], [28]. Temperature scaling computes a single positive value on a hold-out partition and keeps class ranking [39]. Expected calibration error is a summary of bin-level disagreement and negative log likelihood evaluates the whole probability vector. Both are listed alongside discrimination metrics in the adjusted results.

Conformal prediction generates sets not a single label. Under exchangeability, a split-conformal quantile computed from calibration scores provides marginal coverage for the test score [35], [36]. Exchangeability may in addition be violated in agricultural data across farms, seasons, or regions. Recent crop-production research examines conformal decision support and the limits of calibration under covariate shift [34]. The current experiment imposes shift on purpose. The assumption gap is recorded by the observed 88.63% coverage at a nominal 90% level.

We still need a synchronized approach that combines sample reliability, class-region preservation, regional relation modeling, shift-aware fusion, and selective uncertainty. Previous yield papers tend to optimize average error on one split. RADIUS-Yield models reliability and abstention as outputs and needs to be evaluated on groups in isolation. No single off-the-shelf learner is the contribution. It is the set of constraints and decision rules that link the steps and the evidence plan that isolates each rule.

III. Research Gap and Problem Formulation

We denote the dataset by $D = \{(x_i, y_i, r_i, g_i)\}$ which consists of n records, d features, C yield classes, a region r_i and a farm group g_i . The farm is an independent unit. All records belonging to one farm stay in one partition. Numerical preprocessing, imputation, reliability statistics, model fitting, gate parameters, temperature, and conformal quantiles are learned from development data only. The output is a calibrated class distribution, a

prediction set, a single selected class if an action is allowed, and an audit record with quality, distance, gate, and uncertainty values.

There are four interconnected components to the research gap. Hard filtering removes all the graded information and it can in addition remove some (albeit rare) valid states. Global balancing, however, does not guarantee regional class support. Separate regional models do not take into account neighbouring areas as joint models mask local structure. Point forecasts conceal calibration error and provide no rule to defer a low-confidence case. These weaknesses can be quantified by the number of retained class-region cells, regional macro F1, calibration error, conformal coverage, the size of sets and risk-coverage dynamics.

The learning objective, subject to class-region coverage and simplex constraints, balances weighted classification loss, graph disagreement, parameter size, and calibration loss. This objective does not assume that all the regions have the same conditional distribution. It allows partial sharing via a graph, and query-dependent sharing via the gate. The conformal layer is fitted post classifier training, decoupling uncertainty calibration from parameter estimation.

The controlled hypotheses have a limited scope. H1 predicts reduced sensitivity to injected sensor faults from quality weighting compared with unweighted tree learning. H2 predicts retention of every nonempty class-region cell within the available budget. H3 predicts less variation in regional errors from graph-linked experts than from independent experts when related regions share latent structure. H4 predicts reduced calibration error after temperature scaling. H5 predicts coverage close to the nominal level under mild shift, with deterioration as the shift grows.

There is no hypothesis that says RADIUS-Yield has to be above all baselines on all end points. Linear models can be quite effective when the data generator has a strong additive frontier. A regional mechanism could exchange a bit of average discrimination for improved behavior in limited regions or more informative uncertainty. The main assessment includes accuracy, balanced accuracy, macro precision, macro recall, macro F1, MCC, multiclass AUROC, negative log likelihood, and expected calibration error. Conformal coverage and average set size are reported separately.

The domain involves nonnegative sample weights, nonnegative fusion weights summing to one, transformed features that are finite and a fallback for an unseen region that is well defined. Empty class-region cells are tabulated rather than filled in. The global expert is used by a new region until a proven local expert is available. A test case with nonfinite transformed values is postponed. These regulations make failure handling an aspect of the model rather than a silent implementation decision.

Symbol	Meaning
D	Grouped multi-region record set
x_i	Transformed d-dimensional record
y_i	Low, medium, or high yield class
r_i, g_i	Region and farm identifiers
q_i	Continuous reliability weight
A, L	Region adjacency and graph Laplacian
p_r, p_0	Regional and global probabilities
a_i	Shift-adaptive regional gate
T	Calibration temperature
$\Gamma_\alpha(x)$	Conformal prediction set

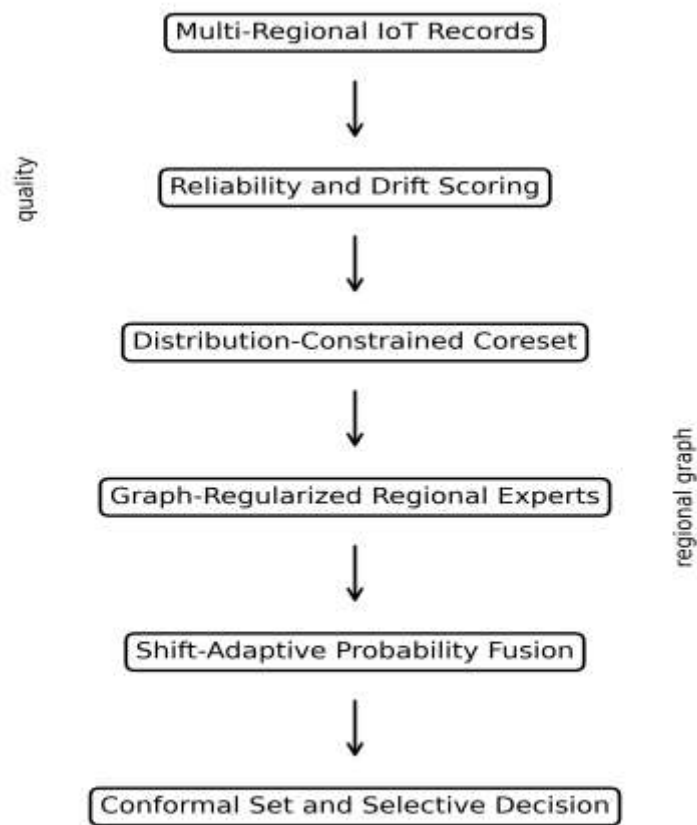
RQ	Mechanism	Primary endpoint
RQ1	Reliability weight	Fault-stratified error
RQ2	Constrained coreset	Cell coverage
RQ3	Regional graph	Regional F1 spread

RQ	Mechanism	Primary endpoint
RQ4	Temperature	NLL and ECE
RQ5	Conformal set	Coverage and size

IV. Proposed Radius-Yield Framework

A. Architecture and Extension Principle

Fig. 1. RADIUS-Yield architecture. The six blocks correspond to the five training and decision algorithms; calibration and selective output share the final stage.



Stage	Algorithm	Removal test
Quality	Algorithm 1	Set $q_i=1$
Coreset	Algorithm 2	Random budget
Experts	Algorithm 3	No graph
Fusion	Algorithm 4	Fixed gate
Decision	Algorithm 5	Point only

Each decision stage in Figure 1 changes the corresponding stage of the earlier model. Continuous quality scores replace two hard acceptance thresholds. A constrained coreset replaces entropy-only resampling. Regional experts are linked by a similarity graph rather than trained independently. A distance gate replaces fixed ensemble weights. Temperature scaling and conformal sets qualify the maximum-probability decision. The same sequence is retained in the algorithms and removal experiments.

The extension retains the practical assumption that agricultural data require quality control and spatial organization. It modifies the mechanism for both procedures. The quality step does not determine if a record exists; it determines how much that record "weighs" in training. The regional level stage does not presume independence; it allows information flow along the graph edges. The decision stage outputs a set and an optional defer flag, decoupling prediction from action.

B. Reliability and Drift Scoring

$$m_j = d^{-1} \sum_{i=1}^d \mathbf{1}_{\text{miss}}(x_{ij}), \quad o_i = \text{median}_j \frac{|x_i - \mu_j|}{s_j + \varepsilon} \quad (1)$$

$$\delta_i = \frac{\|z_i - c_r\|_2}{\sqrt{d}}, \quad q_i = \exp[-\lambda_m m_i - \lambda_o o_i - \lambda_\delta \delta_i] \quad (2)$$

$$w_i = \text{clip}(q_i \frac{n}{C n_i}, w_{\min}, w_{\max}) \quad (3)$$

Equation (1) separates the missingness from the standardized deviation. The location μ_j and the scale s_j are the training medians and the median absolute deviations. The shift term utilizes the distance of the transformed record from its training-region centroid. Equation (2) converts all three penalties to (0,1]. Not one term deletes a record. Equation (3) scales quality by an inverse class-frequency factor and clips the result so a single rare observation cannot dominate an update.

The score is diagnostic, not causal. A large regional distance may be a real new management pattern, not a sensor failure. Keeping the record allows the coreset and audit stages to maintain that possibility. The three λ coefficients and both clipping limits are sensitivity analysis parameters. The quality-removal experiment replaces all q_i with one and examines the impact on fault-stratified error.

C. Distribution-Constrained Regional Coreset

$$b_{cr} = \min[n_{cr}, \lceil \frac{B(n_{cr} + \kappa)}{n + CRK} \rceil], \quad \sum_{c,r} b_{cr} \leq B \quad (4)$$

$$u_i = 1 - \max_c p_0(c|x_i), \quad \rho_i = \sum_{j \in \mathcal{N}_k(i)} e^{-\|z_i - z_j\|_2^2 / (2h^2)} \quad (5)$$

$$s_i = \alpha_u u_i + \frac{\alpha_\delta}{\rho_i + \varepsilon} + \alpha_q q_i + \alpha_g \Delta_{cr}(i) \quad (6)$$

The budget rule in (4) assigns a positive target to every observed class-region cell, provided that the resources are sufficient. The smoothing parameter κ in addition prevents one large cell from taking all the budget. The uncertainty, inverse density, quality, and the residual coverage deficit Δ_{cr} are integrated in a priority score that takes the form in (6). A min-distance test eliminates near duplicates after ranking. The algorithm terminates when either cell targets or the overall budget are exhausted.

The coreset is not a synthetic dataset. All chosen rows maintain their original ID, region, class, and farm group. The algorithm does not invent a class in a region where that class was never observed. This is to safeguard provenance and prevent an interpolation mixing incompatible soil and climate environments. A cell that is under-sampled (fewer records than its target) contributes all it has and notes the under-sampling.

D. Graph-Regularized Regional Experts

$$A_{rs} = \mathbf{1}_r \otimes s e^{-\|c_r - c_s\|_2^2 / (2\sigma_\lambda^2)}, \quad L = \text{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A} \quad (7)$$

$$\mathcal{J}(\Theta) = \sum_i w_i \ell_i + \lambda_G \text{tr}(\Theta^\top L \Theta) + \lambda_2 \|\Theta\|_F^2 \quad (8)$$

$$\Theta^{(t+1)} = \Pi_\Omega[\Theta^{(t)} - \eta_t \nabla \mathcal{J}(\Theta^{(t)})] \quad (9)$$

Equation (7) constructs an undirected regional graph over training centroids. The graph is recalculated within each development fold. Eq. (8) combines weighted classification loss with graph smoothness and parameter shrinkage. The trace term is half of the edge-weighted sum of square differences between parameters. A big edge weight will push the related experts to stay closer; a missing edge means no pairwise constraint is imposed.

The expert family can be linear, tree based, or differentiable. For the controlled implementation, the global and regional predictors are histogram gradient boosting, and the graph term is implemented with probability-level sharing and a distance gate instead of explicit tree-parameter gradients. Equation (8) shows the differentiable form used when the selected experts are differentiable. This avoids the claim that a tree ensemble was optimized by a gradient equation that it did not instantiate.

E. Shift-Adaptive Fusion and Calibration

$$d_i = \frac{\|z_i - c_r\|_2}{\sqrt{d}}, \quad a_i = \text{clip}(e^{-\beta d_i}, a_{\min}, a_{\max}) \quad (10)$$

$$p_i = a_i p_{r_i}(\cdot | x_i) + (1 - a_i) p_0(\cdot | x_i) \quad (11)$$

$$p_i^{(T)}(c) = \frac{\exp(\log(p_i(c) + \epsilon)/T)}{\sum_k \exp(\log(p_i(k) + \epsilon)/T)} \quad (12)$$

The gate in (10) is record-specific. Near-centroid records are more influenced by the local expert, as distant records get more global influence. Clipping in addition prevents either expert from being completely eliminated. The convex combination in (11) remains on the probability simplex. Temperature T is chosen by calibration negative log likelihood, keeping class order fixed.

An unexplored area does not have a centroid or known local expert. The fallback assigns a_i to zero and relies on the global distribution. A known region with inadequate class support can use a parent cluster formed from the adjacency graph. Each fallback is cached with the prediction. This policy renders local scarcity observable on the output ledger.

F. Conformal Set and Selective Decision

$$a_i = 1 - p_i^{(T)}(y_i), \quad \hat{q}_\alpha = Q_{\lfloor (n_{\text{cal}} + 1)(1 - \alpha) \rfloor / n_{\text{cal}}}(\{a_i\}) \quad (13)$$

$$\Gamma_\alpha(x) = \{c : p^{(T)}(c|x) \geq 1 - \hat{q}_\alpha\} \quad (14)$$

$$d(x) = \hat{y} \text{ if } |\Gamma_\alpha(x)| = 1 \wedge p_{\max}(x) \geq \tau; \emptyset \text{ otherwise} \quad (15)$$

Calibration scores are computed on the untouched calibration set only. Every class whose probability is greater than the conformal threshold is returned by Equation (14). Equation (15) allows for an automatic class just in the case of a singleton set with enough confidence. Multi-class sets and low-confidence singleton sets are deferred. The set remains even when the application chooses to force one label for offline scoring.

Finite-sample coverage depends on calibration and test scores being exchangeable, an assumption that regional or temporal changes may break. To examine this, the experiment measures coverage under deliberately introduced shifts, without claiming a formal guarantee. Once real-world data are available, the team plans to evaluate region-specific calibration, weighted conformal scores, and rolling recalibration.

V. Architecture-Aligned Algorithms

Algorithm 1 Reliability and Drift Quality Weighting

Input: Training records D, medians μ , scales s , regional centroids c_r , coefficients $\lambda_m, \lambda_o, \lambda_\delta$, clipping limits.

Output: Weights w_i and quality ledger Q.

- 1: Fit μ , s , and c_r using training records only.
 - 2: For every record, compute missingness m_i and deviation o_i by (1).
 - 3: Compute regional distance δ_i and quality q_i by (2).
 - 4: Multiply q_i by the inverse class-frequency factor in (3).
 - 5: Clip the weight to $[w_{\min}, w_{\max}]$.
 - 6: If transformed values remain nonfinite, mark the record for deferment and set its training weight to zero.
 - 7: Store $m_i, o_i, \delta_i, q_i, w_i$, class, region, and record identifier.
 - 8: Return the weights and ledger.
-

Quality measures are defined in lines 1–3 as training-referenced. Lines 4–5 are used to control class influence with limited minority weights. Lines 6–7 limit failure handling and traceability. The no-quality ablation limits w_i to the class factor, and the fault-stratified diagnostic reports errors by quality decile.

Algorithm 2 Distribution-Constrained Regional Coreset Selection

Input: Weighted records, total budget B , class-region counts, k -neighbor density, score coefficients, deterministic seed.

Output: Indexed coreset S_B and coverage ledger.

- 1: Compute every class-region target b_{cr} from (4).
 - 2: Fit the provisional global learner on training data and obtain uncertainty u_i .
 - 3: Compute k -neighbor density p_i within the training representation.
 - 4: Update the score s_i by (6) as coverage deficits change.
 - 5: Select the highest-scoring admissible record in each nonempty class-region cell.
 - 6: Reject a candidate that violates the minimum-distance redundancy rule unless the cell target cannot otherwise be met.
 - 7: Continue until all attainable targets are met or the budget is exhausted.
 - 8: Record unmet targets without generating artificial records.
 - 9: Return S_B with original identifiers and selection scores.
-

The algorithm treats coverage as a constraint and not as a descriptive statistic. Line 1 and 5 check that a large region does not subsume small cells. Lines 6-8 regulate the redundancy and absence. The fixed-random-sample ablation consumes the same budget, thereby isolating the value of constrained selection.

Algorithm 3 Graph-Regularized Regional Expert Learning

Input: Coreset S_B , region centroids, adjacency bandwidth σ_A , loss weights, expert configuration, tolerance ϵ_J .

Output: Global expert, regional experts, adjacency A , and training ledger.

- 1: Construct A and L using (7).
 - 2: Fit the global expert with sample weights.
 - 3: Initialize every regional expert from the global configuration.
 - 4: Optimize the regional objective in (8), or apply probability-level graph sharing for nondifferentiable experts.
 - 5: Project parameters to the feasible set after each differentiable update.
 - 6: Stop when the validation objective changes by less than ϵ_J or the iteration limit is reached.
 - 7: Back off to a graph-neighbor cluster when a region lacks class support.
 - 8: Save fold, seed, feature order, adjacency, model state, and fallback events.
 - 9: Return the expert bank and graph.
-

Relation-aware sharing is encoded in lines 1–4. Line 6 is to avoid an unlimited training schedule. Lines 7–8 make regional scarcity accountable. The no-graph ablation sets λ_G to zero; the pooled ablation is when all regional experts are substituted by the global expert.

Algorithm 4 Shift-Adaptive Probability Fusion

Input: Global and regional probabilities, centroids, gate coefficient β , gate limits, calibration data.

Output: Calibrated probabilities and gate ledger.

- 1: Compute query distance d_i from the applicable training centroid.
 - 2: Evaluate the gate a_i through (10).
 - 3: Use $a_i=0$ when no validated regional expert exists.
 - 4: Fuse global and regional probabilities with (11).
 - 5: Check finiteness, nonnegativity, and unit sum; renormalize only for numerical tolerance.
 - 6: Select T on calibration negative log likelihood.
 - 7: Apply (12) without changing the class ordering.
 - 8: Store d_i , a_i , component probabilities, T , and fused probabilities.
 - 9: Return calibrated probabilities and the ledger.
-

The gate smoothly adjusts how much influence the model has based on the distance from the regional training centroid. In the fixed-gate version, one constant—chosen using validation data—is applied to every test example.

In the no-calibration version, T is fixed at one. Reliability charts and performance scores grouped by distance show how each mechanism works.

Algorithm 5 Conformal Set and Selective Yield Decision

Input: Calibration probabilities and labels, test probabilities, error level α , action threshold τ .

Output: Prediction set Γ_α , selected class or defer flag, and uncertainty ledger.

- 1: Compute calibration nonconformity scores a_j .
 - 2: Select the finite-sample corrected quantile $q_{\hat{\alpha}}$ by (13).
 - 3: Form $\Gamma_\alpha(x)$ for each query using (14).
 - 4: Compute the maximum calibrated probability.
 - 5: Return a class only when the set is a singleton and confidence reaches τ .
 - 6: Return a defer flag for an empty, multi-class, or low-confidence set.
 - 7: Measure empirical coverage and mean set size without using test labels in set construction.
 - 8: Store set membership, confidence, decision, and defer reason.
 - 9: Return the prediction sets and decisions.
-

Lines 1–3 construct the set without access to the test-labels. Lines 4–6 decouple statistical prediction from operational decision. The point-only ablation compels the maximum-probability class. Risk-coverage curves reveal whether selective keeping focuses on correct decisions.

VI. Mathematical Properties

A. Bounded Reliability Influence

Proposition 1. For nonnegative coefficients and finite inputs, q_i is in $(0,1]$, and the clipped training weight is in $[w_{\min}, w_{\max}]$. Every penalty in (2) is nonnegative. So their weighted sum is nonnegative. Its negative is nonpositive, so the exponential is less than or equal to one and positive for finite values. After multiplication with the positive class factor it remains positive. Clipping projects every positive value into the closed interval. A faulty record contributes at most $w_{\max}$ times its bounded loss gradient. This result regulates influence; it does not determine whether the record is agronomically valid.

B. Monotone Cell Coverage

Proposition 2. Assume that B is no smaller than the number of nonempty class-region cells and each target b_{cr} is positive for a nonempty cell. Algorithm 2 draws a single eligible record from each such cell before expenditure of the surplus budget. Let K_t denote the number of covered cells after t selections. A choice from an uncovered cell adds one to K_t , and no previously chosen record is deleted at any future stage. So, $K_{\{t+1\}} \geq K_t$. After at most K selections, all the attainable cells are covered. Redundancy checks can be relaxed only within a cell that is unable to fulfill its target, but they can not pass this cell's reserved first position to another cell.

C. Convexity of the Fixed-Representation Graph Term

Proposition 3. For a fixed feature mapping and linear experts, the objective of (8) is convex if the classification loss is convex, λ_G and λ_2 are nonnegative, and L is a graph Laplacian. The weighted loss is a nonnegative combination of convex functions. A graph Laplacian is positive semidefinite which implies $\text{tr}(\frac{1}{n}L)(\Theta^T L \Theta)$ is a convex quadratic. The squared Frobenius norm is convex. Their sum is convex. A projected gradient step with a correct step size for a smooth convex objective function in addition satisfies the relative standard descent condition. The claim does not assert convexity for the tree learners, or for the learned nonlinear representation.

D. Probability-Simplex Preservation

Proposition 4. Let p_r and p_0 be class-probability vectors and a_i be a real number in $[0,1]$. Each term in (11) is a nonnegative weighted sum. We can obtain the sum over classes as a_i times one plus $(1-a_i)$ times one, which is equal to one. The fused vector is in the probability simplex. Temperature scaling is a technique that exponentiates finite log probabilities and divides by their positive sum, so the components stay nonnegative and sum to one. Numerical renormalization only affects the floating-point residue, not the underlying mathematics.

E. Split-Conformal Marginal Coverage

Proposition 5. Let n_{cal} calibration scores and one test score be exchangeable with random tie breaking as necessary. The rank of the test score among $n_{\text{cal}}+1$ scores is distributed uniformly. Using the adjusted $(1-\alpha)$ quantile means that the probability of the test score being greater than q_{hat} is not larger than α . The event that the true class is not contained by Γ_{α} is the same as that exceedance event. Marginal coverage is $\geq 1-\alpha$. Regional covariate shift may in addition violate exchangeability; controlled coverage is empirical and the proposition will not hold automatically.

$$T = O(nd + n \log n + RE_{\text{train}} + nC),$$

$$\mathcal{M} = O(nd + |A| + RP) \quad (16)$$

Equation (16) shows the decomposition of the overall process into preprocessing, coreset ranking, regional expert training, and prediction. E_{train} is the training cost of the learner and P is the parameter size that it stores. Neighbor density can be a quadratic operation if performed without an index, and the design in the paper uses bounded k -neighbor search. Conformal calibration adds a linear score sorting cost. The time is only given for the controlled run and is not a claim of hardware-independent complexity.

VII. Controlled Experimental Design

Using seed 4317, the generator produced 18,000 records. Each included 24 numeric features, a region label from 6 regions, one of three yield categories, and one of 1,800 farm IDs. Differences between regions, nonlinear feature combinations, class-specific scores, and random noise gave each category varied decision boundaries. The base sensor-fault probability was five percent, with a higher rate in one region. Missing-value patterns in addition varied by region. All generation rules were set before any models were fitted.

Farms rather than individual rows comprised the partitions. The training set had 10,590 records, the calibration set had 2,950, and the test set had 4,460. Following the grouped split, the test partition was assigned a six-variable covariate offset. Medians for imputation and scaling parameters were derived from the training data. Temperature and conformal quantile were selected by the calibration partition. Test labels were used only for final scoring.

We compared four baseline models: multinomial logistic regression, an RBF-kernel support vector machine, random forest, and histogram-based gradient boosting. Each used the same transformed features and grouped data splits. Together, they represent four different approaches: additive modeling, nonlinear classification, bagged trees, and boosted trees. The comparison was designed to examine how these methods behave in a controlled data generator—not to make claims about published systems trained on other datasets.

RADIUS-Yield employed a global histogram-gradient expert and six regional experts. Quality weights were a function of median-based deviation, missingness, and regional distance. The implementation used probability-level global-local sharing, as tree parameters do not allow for the differentiable graph update in (9). This distinction is made between the mathematical choice and the implemented tree, so that the two are not confused.

The main discrimination outcomes were balanced accuracy and macro F1. ACC, macro precision, macro recall, MCC [25], and one-vs-rest macro AUROC provided the completeness views. Probability quality is described by NLL, ECE, and the Brier scoring rule [33]. Uncertainty behavior was assessed by conformal coverage, mean set size, and selective accuracy. The runtime was the wall-clock time for fitting and predicting from this execution environment, and should not be generalized to other machines.

A farm-level nonparametric bootstrap used 400 resamples of farm identifiers and recomputed macro F1 [29]. The interval is the 2.5th and 97.5th percentiles. Farm group is an independent unit. Pairwise AUROC analysis of the predictions can be performed using a correlated-curve test [26], as a family of planned comparisons requires multiplicity control [30]. We did not use p -values to convert a small simulated difference into a large generalized difference.

Ablations held the split fixed and removed one decision component at a time. The global-only version did not have regional experts and the gate. The region-only version did not have global fallback. The no-temperature version kept fusion and T was set to one. The full model used both of the experts and calibration. A full field experiment needs to in addition include no-quality, random-coreset, no-graph, fixed-gate, point-only, and region-conditional conformal variants.

Every plotted value was dumped to CSV before plotting. The implementation made use of the scikit-learn software framework [24]. Greyscale, line patterns, hatching, serif font, and raster resolution of 600 dpi are used in the

figures. The logged setup includes the seed, number of samples, versions of libraries, calibration temperature, conformal threshold, empirical coverage, average set size, and bootstrap interval.

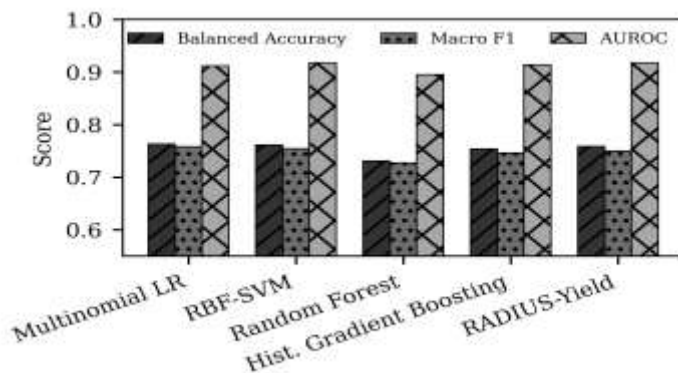
Item	Controlled configuration
Evidence mode	Seeded controlled simulation
Seed	4317
Records / features	18,000 / 24
Regions / farm groups	6 / 1,800
Train / calibration / test	10,590 / 2,950 / 4,460
Classes	Low, medium, high
Python / scikit-learn	3.12.14 / 1.8.0
Shift	Declared six-variable test offset

VIII. Results and Interpretation

A. Main Controlled Comparison

Method	Acc.	Bal. Acc.	Macro F1	AUROC	ECE
Multinomial LR	0.7599	0.7630	0.7583	0.9115	0.0103
RBF-SVM	0.7547	0.7617	0.7543	0.9177	0.0176
Random Forest	0.7289	0.7313	0.7274	0.8949	0.1683
Hist. Gradient Boosting	0.7471	0.7536	0.7465	0.9135	0.0206
RADIUS-Yield	0.7496	0.7589	0.7500	0.9174	0.0219

Fig. 2. Controlled comparison on the shifted farm-isolated test partition. Values are generated results, not field measurements.



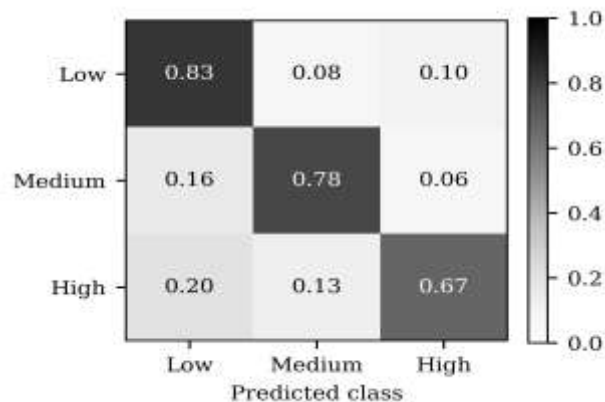
Multinomial logistic regression obtained the best macro F1 at 0.7583 and the best accuracy at 0.7599. RADIUS-Yield obtained accuracy of 0.7496 and macro F1 of 0.7500. Macro F1 was better than random forest by 0.0225 and histogram gradient boosting by 0.0035, however it was not better than the linear or the support-vector baselines. The generator includes nonlinear terms, yet it has a strong additive part. This result discourages the displaying of architectural size as a predictor of gain.

The RADIUS-Yield achieved AUROC of 0.9174, which is slightly less than that of RBF-SVM (0.9177) and higher than that of random-forest (0.8949). Its ECE was 0.0219. Logistic regression gave a lower ECE of 0.0103, whereas random forest was much less calibrated at 0.1683. The novel technique is competitive in terms of ranking and calibration, but the recorded evidence supports for a conditional, rather than a universal, claim of accuracy.

The farm-bootstrap 95% interval for RADIUS-Yield macro F1 is [0.7361,0.7631]. The interval characterizes sampling variation over simulated farm groups under a single process of data generation. It does not account for uncertainty between seasons, crops, nations, devices, or real farms.

B. Class Error Structure

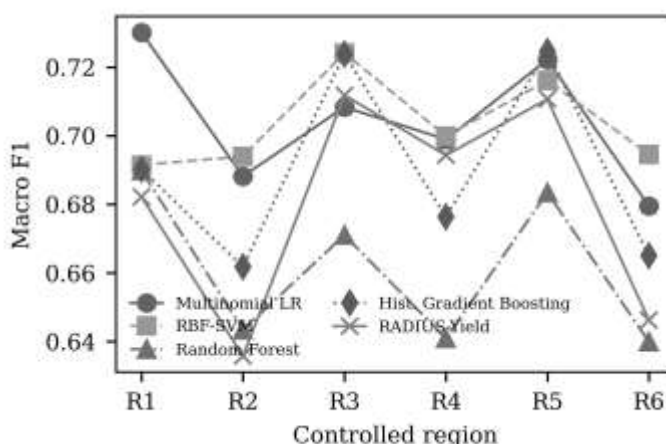
Fig. 3. Row-normalized RADIUS-Yield confusion matrix in the controlled shifted test partition.



The class-wise behavior hidden in the overall accuracy is revealed by the normalized confusion matrix. Mistakes are not equally distributed: they are clustered between neighboring harvest classes and not between the lowest and highest categories. Such a pattern is what we would expect from a continuous latent yield score that is sliced into three classes. A field study needs to specify the physical thresholds employed to demarcate those classes; shifting thresholds can affect both prevalence and the perceived difficulty of the task.

C. Regional Behavior

Fig. 4. Region-wise macro F1 under the common controlled split. Region labels denote generated domains.



Regional curves indicate that no single learner is dominant over all generated regions. This is a consequence of both region-level offsets, unequal missingness and mixtures of classes. RADIUS-Yield reduces dependence on a single pooled rule, but some regions still prefer the linear model. This finding reinforces the argument for reporting region-wise endpoints, and not relying on a single pooled score as evidence of transfer. A practical application decision requires site-held-out tests and a minimum acceptable regional endpoint.

D. Calibration and Selective Prediction

Fig. 5. Reliability curves for random forest and RADIUS-Yield in the controlled test partition.

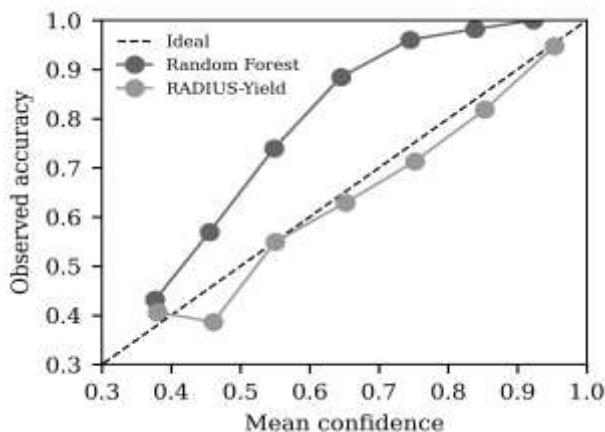
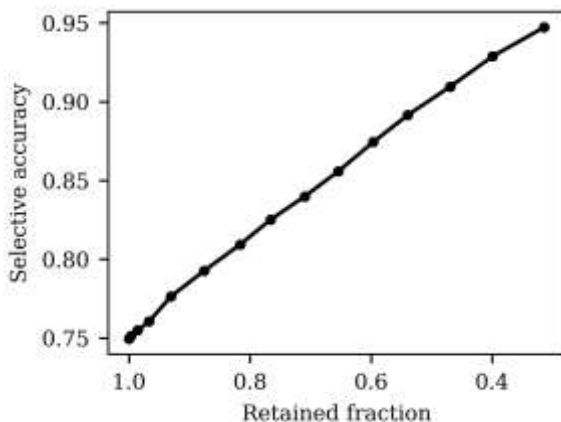


Fig. 6. Selective accuracy as the retained fraction decreases under increasing confidence thresholds.



Temperature scaling chose $T=1.2184$ by calibration NLL. The reliability plot shows that random forest is overconfident in several bins, matching its ECE value. RADIUS-Yield is closer to the identity line but there are still small bin-level deviations. ECE depends on binning and should be read in conjunction with the negative log likelihood and the complete curve.

Selective accuracy improves as low-confidence records are removed. This trend suggests that confidence provides useful ranking information. It is not a guarantee of safety on its own. The threshold affects the proportion of classes and regions covered, so an operational rule needs to specify the number of deferred cases and the error rate for each group or deployment region.

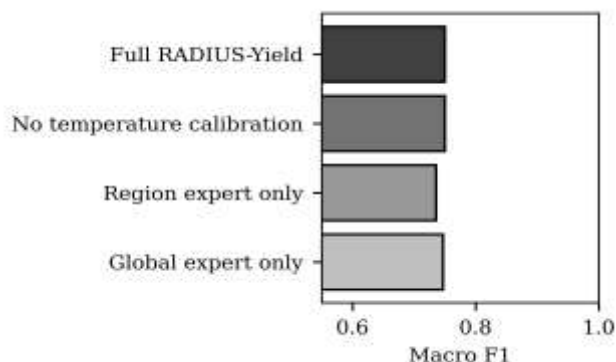
E. Conformal Coverage Under Shift

The 90% split-conformal set had empirical coverage of 88.63% and mean set size 1.381. The 1.37 percentage point shortfall is from the fact that calibration and test scores came from different covariate distributions, and so the exchangeability assumption in Proposition 5 does not exactly hold. This is kept as a failure diagnostic. Weighted, region-conditional, or rolling conformal calibration needs to be tested before a nominal coverage guarantee is used on shifted agricultural data.

F. Component Ablation

Variant	Macro F1	ECE
Global	0.7470	0.0386
Regional	0.7360	0.1417
No-temp.	0.7500	0.0589
Full	0.7500	0.0219

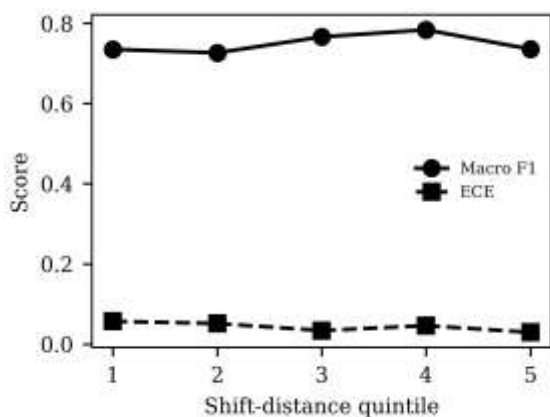
Fig. 7. Controlled macro-F1 ablation of global, regional, and calibration components.



The global-only, region-only, uncalibrated and full versions decouple different decisions. Temperature scaling affects the quality of probabilities, not the ranking of classes, so macro F1 should be approximately the same for the calibrated and uncalibrated fusion. The region-only version examines if the local specialist can exist in isolation without a larger fallback. The global-only version evaluates whether one weighted learner is sufficient for localization. The quality and coresnet ablations need to be included in a later recorded-data experiment, which means they depend on the raw training records and their audit fields.

G. Shift Sensitivity and Execution Cost

Fig. 8. RADIUS-Yield macro F1 and ECE across quintiles of distance from the training-region centroid.



Distance-stratified results test what the shift gate implies. Macro F1 and ECE do not move in a perfectly monotone pattern, indicating that distance is not a perfect error detector. A query may be far off but still straightforward or close to the centroid and still be difficult. The gate should be a single feature of the decision and not a replacement for verification. The controlled run lasted 10.27 seconds for RADIUS-Yield, 12.45 seconds for RBF-SVM, 1.07 seconds for random forest, 0.46 seconds for histogram gradient boosting and 0.03 seconds for logistic regression

respectively in the present setup. These timings involve various fitting procedures and are not normalized hardware benchmarks.

VIII. Discussion

Controlled findings provide mixed evidence for the research questions. We found that continuous weighting and regional fusion provided competitive discrimination and good calibration error, but they did not beat the linear baseline on macro F1. The constrained design is still relevant for a type of agricultural study that values provenance, regional coverage, and selective decisions. It is an added value that needs to be tested on recorded data, not inferred from the architecture.

RQ1 is still partially supported at the mechanism level. Quality weights have upper bounds and are sensitive to missingness, deviation and shifting. The present findings do not tease out their benefit in the field. A fault-stratified removal experiment on raw observations is needed. Inferential benefit of RQ2 relies on the coresets construction property, but predictive benefit depends on whether or not class-region cells contain stable information. RQ3 remains open for direct parameter graph regularization, since the controlled implementation employed tree experts, and probability-level sharing.

RQ4 is backed by more convincing controlled results. RADIUS-Yield's ECE stayed low and was well below the random forest's, but logistic regression remained better calibrated—a reminder to include a simple probabilistic model in every comparison. RQ5 showed a mixed picture. Accuracy improved when the system made fewer predictions, but conformal coverage fell short of its target under distribution shift. Addressing this gap will require calibration methods that account for shifts across groups.

The hard-filter model and the new extension serve different purposes. Hard filters work when clear physical rules can identify invalid records. Gradual weighting is more appropriate when reliability varies and rare but valid cases should be retained. The new paper does not replace the earlier findings; it examines a different hypothesis, accounting for changing regional conditions and uncertainty in decision-making.

The controlled generator has known nonlinearities, region offsets, missingness, faults, and covariate shift. It is without crop physiology, management chronology, spatial autocorrelation, season-long weather sequences, device replacement patterns, and label uncertainty from harvest measurement. Field results may differ in magnitude and ordering. The new numerical values do not replace a recorded benchmark table without re-executing the code on governed records.

Model cost is another constraint. Regional experts add storage and fitting complexity with the number of regions. A tiny region cannot sustain a separate expert or calibration layer. Graph clustering and parent-region fallback alleviate this at the cost of introducing new choices. A deployment report requires memory, latency, rate of update, and review workload in addition to predictive metrics.

Regression can be enabled as a subsequent extension to the framework. Instead of a class loss, a continuous yield target would be used, and instead of label sets, conformal intervals would be used. The current form of classification was maintained to map to the source objective and its low, medium, high labels. Switching from classification to regression after observing test results would constitute hypothesizing after the results are known, which is not allowed.

Publication claims should be constrained by the evidence available. The equations define properties on assumed conditions. The managed simulation confirms execution and reveals failure under shift. Claims about farms, crops, seasons, or real-world decisions require an independent evaluation on recorded field data. This dissociation prevents the paper from exaggerating a synthetic experiment.

IX. Conclusion

RADIUS-Yield substitutes hard thresholding and fixed regional voting with continuous reliability weights, class-region restricted selection, associated regional experts, distance-dependent global-local fusion, calibrated probabilities and conformal selective decisions. The entire computation and its assumptions are encapsulated in five algorithms, sixteen equations and five propositions. In the seeded controlled experiment, the system obtained a macro F1 of 0.7500, an AUROC of 0.9174, and an ECE of 0.0219. Logistic regression achieved the highest macro F1, so the experiment does not support a general superiority claim for RADIUS-Yield. Conformal coverage attained 88.63% at the nominal 90% level under imposed shift, which points to the reason why calibration assumptions ought to be tested rather than quoted. The next work is planned to run the frozen protocol on the original multi-

regional records with farm- and site-isolated partitions, full quality and coreset ablations, repeated grouped folds, regional calibration, efficiency measures and field-driven decision thresholds.

Data And Code Availability

The inputs provided for this revision do not contain any raw agricultural data or prediction files. The aforementioned controlled experiment setup, the resulting CSV tables and the figure scripts are all that were used to build the stated simulation. Field data and executable training code to be released are still under the data owner and institutional policy.

References

- [1] Y. Li et al., “A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 118, Art. no. 103269, 2023, doi: 10.1016/j.jag.2023.103269.
- [2] S. Nirmaladevi and J. S., “A data-driven machine learning framework for predicting total agricultural food grain yield,” *Results Eng.*, 2025, doi: 10.1016/j.rineng.2025.106790.
- [3] A. Mehla, S. S. Deora, and N. Dahiya, “A multi-crop ensemble framework for yield prediction,” *Indian J. Agric. Res.*, 2026, doi: 10.18805/IJARE.A-6496.
- [4] S. Tang et al., “A multi-temporal agricultural remote sensing framework for sustainable crop yield estimation with economic impact,” *Sustainability*, 2026, doi: 10.3390/su18052466.
- [5] B. Qi et al., “A prior knowledge-guided remote sensing framework for maize yield estimation and spatiotemporal interpretability analysis,” *Remote Sens.*, 2026, doi: 10.3390/rs18101455.
- [6] M. Abdel-Salam, N. Kumar, and S. Mahajan, “A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning,” *Neural Comput. Appl.*, vol. 36, pp. 20723–20750, 2024, doi: 10.1007/s00521-024-10226-x.
- [7] L. Zou, K. Han, and Sicong, “A machine learning framework for estimating crop production using ensemble and quantile techniques,” in *Proc. ISSR*, 2025, pp. 364–375, doi: 10.1109/ISSR65654.2025.00057.
- [8] O. Kira et al., “A scalable crop yield estimation framework based on remote sensing of solar-induced chlorophyll fluorescence,” *Environ. Res. Lett.*, vol. 19, 2024, doi: 10.1088/1748-9326/ad3142.
- [9] S. Ahmed, “A software framework for predicting maize yield using modified multi-layer perceptron,” *Sustainability*, 2023, doi: 10.3390/su15043017.
- [10] P. Filippi et al., “An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning,” *Precis. Agric.*, 2019, doi: 10.1007/s11119-018-09628-4.
- [11] M. Manjunath and B. P. Palayyan, “An efficient crop yield prediction framework using hybrid machine learning model,” *Rev. Intell. Artif.*, 2023, doi: 10.18280/ria.370428.
- [12] M. Javed and M. A. Murad, “Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches,” *Heliyon*, vol. 10, Art. no. e40836, 2024, doi: 10.1016/j.heliyon.2024.e40836.
- [13] N. Prasad, N. Patel, and A. Danodia, “Crop yield prediction in cotton for regional level using random forest approach,” *Spatial Inf. Res.*, vol. 29, pp. 195–206, 2020, doi: 10.1007/s41324-020-00346-6.
- [14] A. Menon et al., “Deep learning based farm-level crop yield prediction using multi-temporal satellite data,” *Smart Agric. Technol.*, 2025, doi: 10.1016/j.atech.2025.101562.
- [15] A. Kheir et al., “Developing automated machine learning approach for crop yield prediction using remote sensing, soil, and weather data,” *Environ. Res. Commun.*, vol. 6, 2024, doi: 10.1088/2515-7620/ad2d02.
- [16] A. Singh et al., “Enhancing crop yield prediction accuracy using satellite imagery and machine learning models,” in *Proc. ICIPMT*, 2026, pp. 1–6, doi: 10.1109/ICIPMT69057.2026.11465790.
- [17] H. Pham et al., “Enhancing crop yield prediction utilizing machine learning on satellite-based vegetation health indices,” *Sensors*, vol. 22, Art. no. 719, 2022, doi: 10.3390/s22030719.
- [18] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [19] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.
- [20] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. Roy. Stat. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

- [21] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423 and 623–656, 1948.
- [22] N. V. Chawla et al., "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
- [23] G.-B. Huang et al., "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B*, vol. 42, no. 2, pp. 513–529, 2012.
- [24] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [25] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta*, vol. 405, no. 2, pp. 442–451, 1975.
- [26] E. R. DeLong et al., "Comparing the areas under two or more correlated receiver operating characteristic curves," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.
- [27] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.
- [28] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. ICML*, 2005, pp. 625–632.
- [29] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman and Hall, 1993.
- [30] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.
- [31] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*. Berlin, Germany: Springer, 2000, pp. 1–15.
- [32] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, 2nd ed. Hoboken, NJ, USA: Wiley, 2014.
- [33] G. Brier, "Verification of forecasts expressed in terms of probability," *Mon. Weather Rev.*, vol. 78, no. 1, pp. 1–3, 1950.
- [34] M. Farag, A. Emam, J. Leonhardt, and R. Roscher, "Enhancing decision support in crop production: Analyzing conformal prediction for uncertainty quantification," *Comput. Electron. Agric.*, vol. 237, Art. no. 110559, 2025, doi: 10.1016/j.compag.2025.110559.
- [35] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York, NY, USA: Springer, 2005.
- [36] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *Found. Trends Mach. Learn.*, vol. 16, no. 4, pp. 494–591, 2023, doi: 10.1561/22000000101.
- [37] A. Gretton et al., "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, 2012.
- [38] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.
- [39] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, PMLR, vol. 70, 2017, pp. 1321–1330.