International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(14s) 2026 pp. 869-901
https://doi.org/10.70917/ijcisim-2026-4287
Article

AgriFuse-SSL: Region-Adaptive Weak-Strong Filtering and Confi-dence-Calibrated Semi-Supervised Learning for IoT Crop-Yield De-cision Support

Phanikanth Chintamaneni[1], G. Krishna Mohan[2], Kodukula Subrahmanyam[3]

[1] Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.
[2] Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.
[3] Professor, Department of Computer Science and Engineering, Dr. RVR NRI Institute of Technology (Deemed to be University), Pothavarappadu, Agiripalli – 521212, Andhra Pradesh, India.
Email: smkodukula@yahoo.com

Abstract: — Agricultural Internet-of-Things deployments produce large, heterogeneous streams whose labels are sparse, whose sensor quality varies across seasons and regions, and whose class distributions can drift with crop and management conditions. These properties limit the reliability of crop-yield decision systems trained either on a comparatively small labeled regional dataset or on a larger but weakly supervised sensing pool. This paper proposes AgriFuse-SSL, a unified region-adaptive framework that combines the complementary mechanisms of two existing project models: the weak-strong filtering, informative-sample selection, and reliability-weighted ensemble model; and the pseudo-label generation, confidence-based selection, expanded sensing schema, and optimized semi-supervised convolutional predictor of large scale. The proposed framework first locks a 53-predictor schema, isolates independent farm-field records by record hash and time-region group, and applies weak and strong reliability gates that jointly consider missingness, local density, physical-range violations, temporal inconsistency, and model disagreement. It then forms density-calibrated pseudo-labels, corrects their regional prior, selects features using cross-region stability, and learns a compact feature-token encoder that combines one-dimensional convolution with tabular self-attention. A reliability-weighted regional ensemble produces low, medium, or high yield classes together with calibrated confidence and an accept-or-refer decision. The source specification contains 1,000,000 Ensemble model records with 35 predictors and 3,000,000 Scalable model records with an expanded 53-predictor schema. On a 100,000-record simulated test set, AgriFuse-SSL achieved 0.990 accuracy, 0.990 macro-precision, 0.990 macro-recall, 0.990 macro-F1, and 0.996 macro-AUC. The mechanism ablation increased accuracy from 0.971 for the supervised anchor to 0.990 for the complete model, while the proposed compact implementation reduced simulated batch latency to 72 ms. Region-stratified, calibration, risk-coverage, feature-stability, and uncertainty analyses show that proposed model more effective than treating filtering, pseudo-labeling, and prediction as independent stages.
Keywords: smart agriculture, agricultural Internet of Things, crop-yield prediction, semi-supervised learning, pseudo-labeling, weak-strong filtering, tabular attention, regional generalization, uncertainty calibration, ensemble learning

1. INTRODUCTION

Agricultural production is increasingly observed through networks of soil probes, weather stations, irrigation controllers, remote-sensing indices, farm-management records, and mobile decision-support systems. This transition has created an opportunity to move from coarse seasonal advice toward field- and time-specific decisions. Precision agriculture is commonly understood as the coordinated use of sensing, analytics, and intervention to improve resource efficiency while maintaining or increasing production. Earlier smart-farming studies emphasized connectivity and data acquisition, whereas current systems are expected to transform heterogeneous measurements into predictions that remain dependable when sensors fail, seasons change, or a model is deployed outside the dominant training region

[1]–[6]. Crop-yield classification is therefore not merely a pattern-recognition task. It is an evidence-integration problem in which the input distribution, label quality, spatial support, and decision confidence must be handled explicitly.

The technical difficulty begins with the data-generating process. Temperature, rainfall, solar radiation, soil moisture, nutrient measurements, irrigation activity, normalized difference vegetation index, crop age, pest pressure, and categorical descriptors are produced at different temporal frequencies and measurement scales. A soil-moisture probe can become biased after installation, rainfall records may be accumulated over a different window than canopy temperature, and categorical values such as soil type or agro-climatic zone can be inconsistently encoded. Missingness is rarely independent of agricultural conditions: communication failure may be more frequent during storms, a dry-field station may be visited less often, and expensive laboratory nutrients may be measured only for selected farms. A classifier that treats these omissions as random can learn the acquisition process instead of the agronomic response. Reviews of machine learning in agriculture repeatedly identify data quality, interoperability, transferability, and limited ground truth as central barriers to operational adoption [7]–[10].

The second difficulty is the mismatch between record volume and trustworthy labels. Sensor networks can generate millions of rows, yet reliable yield labels require harvest measurement, field-boundary reconciliation, crop identification, and sometimes manual adjudication. The resulting labeled set is typically small relative to the unlabeled or weakly labeled sensing pool. Supervised learning discards the majority of available information; naive self-training assigns a label to every unlabeled record and can amplify early errors. Semi-supervised learning offers a principled middle position by combining supervised loss with consistency or pseudo-label objectives [29]–[37]. Nevertheless, methods developed for natural images cannot be transferred mechanically to structured agricultural data. Image augmentations such as cropping and color jitter have clear invariances, while aggressive changes to rainfall, soil pH, or crop age can violate physics or agronomy. Weak and strong perturbations must therefore be feature-aware, range-constrained, region-sensitive, and traceable to plausible sensor noise.

The third difficulty is spatial and seasonal heterogeneity. A relationship learned in an irrigated delta may not hold in dryland or highland agriculture. Soil texture changes the effect of rainfall on plant-available water; crop calendars alter the meaning of accumulated heat; irrigation source affects both reliability and management response. A single global model can achieve a high average score while failing in a smaller region. Conversely, independent regional models may be statistically inefficient and impossible to maintain. Region-aware learning should share stable structure while allowing local corrections. It should also prevent a large region from dominating pseudo-label generation. This requirement motivates a feature-selection criterion that rewards relevance and predictive contribution but penalizes instability across region, crop-season, and resampled training subsets.

The fourth difficulty is confidence. A probability vector is not automatically a reliable measure of correctness. Modern neural networks can be overconfident, especially under distribution shift [38]–[42]. Agricultural decision support needs more than a class label because the consequence of a wrong recommendation depends on context. A low-confidence prediction caused by an unseen sensor pattern should be referred for agronomic review instead of silently converted into an action. Calibration, ensemble disagreement, and selective prediction make this behavior measurable. Risk-coverage analysis asks how error changes as the system accepts only increasingly reliable cases. Conformal and calibrated prediction methods further demonstrate that uncertainty should be treated as an output with an explicit operational meaning rather than a decorative score [43], [44].

The project presentation supplied for this study defines four large Andhra Pradesh agricultural IoT models. Ensemble model and Scalable model are especially complementary. Ensemble model contains 1,000,000 records, 35 predictors, five categorical variables, a three-class YieldClass target, and a methodology based on weak and strong filtering followed by informative-sample selection and ensemble classification. Its reported comparison includes kernel extreme learning machine (KELM), kernel logistic regression (KLR), and classification support vector machine (CSVM), with the Ensemble model proposed method reaching 0.975 accuracy, 0.977 precision, 0.974 recall, 0.976 F1, and 0.979 AUC. Scalable model expands the source specification to 3,000,000 records and adds weather, soil, crop-health, geography, irrigation, and agro-climatic variables. Its methodology trains an initial convolutional network on labeled data, generates pseudo-labels for unlabeled records, retains high-confidence examples, ranks features, retrains an extended convolutional predictor, and selects the minimum-MAE checkpoint. The Scalable model proposed method reports 0.984 accuracy, 0.982 precision, 0.983 recall, 0.982 F1, and 0.983 AUC.

Although both models are useful, their separation leaves several unresolved questions. Ensemble model does not fully exploit the larger unlabeled pool or the expanded sensing schema. Scalable model does not explicitly integrate Ensemble model's dual reliability filtering, regional ensemble, and informative-sample mechanism before pseudo-labeling. Neither specification gives a unified rule for preventing region-majority pseudo-labels from overwhelming

minority regions. Their feature-selection stages do not quantify cross-region stability, and their final probabilities are not connected to calibration or referral. In addition, a dataset-summary slide lists 50 Scalable model features while the detailed experimental slide reports 53 features, comprising 46 numerical and seven categorical predictors. The present paper resolves this mismatch by locking the expanded 53-predictor research schema obtained by adding 18 variables to the 35-predictor Ensemble model schema. Any earlier 50-column execution snapshot is treated as a preliminary export, while the 53-predictor dictionary is used for the combined model.

The central hypothesis is that filtering and semi-supervision should not be independent modules. If pseudo-labels are generated before sensor-quality and density checks, the model learns from unreliable records. If filtering is performed without uncertainty, rare but valid regional samples may be removed as outliers. AgriFuse-SSL therefore uses a coupled design. A weak gate removes clear schema and range failures while preserving coverage. A strong gate adds local-density, temporal, and model-disagreement evidence. The retained records support density-calibrated pseudo-labeling, and pseudo-label reliability feeds back into sample weights. Region-stable feature selection then chooses variables that are predictive without depending excessively on one geographic stratum. A compact convolutional feature extractor captures local interactions in the ordered sensor vector, while feature-token attention models nonlocal relationships between weather, soil, crop-health, and management variables. The final regional ensemble is weighted by validation performance, calibration error, and domain similarity.

This paper makes the following contributions.

1. It introduces AgriFuse-SSL, a new combined architecture that unifies Ensemble model's regional weak-strong filtering and ensemble mechanisms with Scalable model's expanded schema and confidence-controlled semi-supervised learning.
2. It formulates an adaptive reliability score that combines physical-range validity, robust missingness, local density, temporal consistency, and predictive disagreement without treating every rare regional sample as noise.
3. It proposes density-calibrated, region-prior-corrected pseudo-labeling. The acceptance threshold is adaptive to class and region support, while a bounded importance weight limits the influence of uncertain pseudo-labels.
4. It develops region-stable feature selection based on relevance, conditional information, bootstrap stability, and between-region dispersion. The selected subset is interpretable and is directly aligned with the expanded 53-feature schema.

The scope of the empirical claims is deliberately bounded. The presentation specifies source CSV endpoints, but the corresponding raw files were not available to the execution environment used to prepare this manuscript. Consequently, the paper does not claim a completed field evaluation on four million observed farm records. It reports presentation-provided Ensemble model and Scalable model aggregates separately from a deterministic schema-calibrated simulation designed to test the internal behavior of the combined framework. The simulation is useful for checking metric consistency, algorithm sequencing, visualization, sensitivity, and statistical procedures; it must be replaced or confirmed by an independent farm-field-season rerun on the source CSV files before claims of field generalization are made.

The remainder of the paper is organized as follows. Section II reviews agricultural IoT prediction, conventional ensembles, tabular deep learning, semi-supervised learning, uncertainty, and regional robustness. Section III defines the research problem and unified data model. Section IV presents AgriFuse-SSL, its algorithms, proofs, complexity, and implementation safeguards. Section V specifies the experimental protocol. Section VI reports controlled results and interpretations. Section VII discusses implications, limitations, and reproducibility. Section VIII concludes the paper and identifies the validation sequence required for deployment.

2. RELATED WORK

A. Agricultural IoT and Precision-Farming Analytics

The development of smart agriculture has been driven by the decreasing cost of sensors, wireless connectivity, cloud storage, and learning systems. Wolfert *et al.* described big-data smart farming as a cyber-physical information chain connecting sensing, analytics, and farm management [1]. Ayaz *et al.* examined IoT architectures for smart agriculture and emphasized communication, energy, scalability, and security constraints [5]. Khanna and Kaur traced the evolution of IoT-enabled agricultural monitoring and highlighted the need to integrate heterogeneous

devices with decision processes [6]. These studies establish that prediction accuracy alone is insufficient; a useful design must respect acquisition constraints and the temporal-spatial structure of farm data.

Machine-learning reviews show rapid growth in classification, regression, disease detection, irrigation prediction, and crop recommendation [2]–[4], [7]. Liakos *et al.* organized agricultural machine-learning applications across crop, livestock, water, and soil management [3]. Kamilaris and Prenafeta-Boldú reviewed deep-learning applications and noted the gap between promising experimental results and operational generalization [2]. Jha *et al.* surveyed automation with artificial intelligence, IoT, robotics, and remote sensing [7]. Sishodia *et al.* reviewed remote-sensing technologies for precision agriculture, including spectral and spatial information that complement ground sensors [8]. Collectively, these works motivate the expanded Scalable model schema, which adds weather, crop-health, geographic, irrigation, and agro-climatic variables to Ensemble model.

Agricultural IoT data are commonly treated as an ordinary tabular matrix after acquisition. That simplification can hide temporal leakage and spatial dependence. Random row splitting may place measurements from the same farm, station, or season in both training and test sets. A classifier can then identify site signatures rather than learn yield mechanisms. The present study uses record hashing and grouped partitions by region and time. This choice is consistent with the broader principle that the unit of inference must match the unit of deployment. In the absence of a stable farm identifier in the source specification, a composite group key is formed from region, district, crop, season, time window, and quantized geographic coordinates. The key is not a substitute for a true farm identifier, but it is safer than an unconstrained random split.

B. Crop-Yield Prediction

Crop-yield prediction has been investigated with process models, remote sensing, classical machine learning, and deep networks. The systematic review by van Klompenburg *et al.* found that temperature, rainfall, and soil variables are among the most common predictors, while artificial neural networks, support vector machines, and tree ensembles are frequent modeling choices [9]. Crane-Droesch demonstrated that semiparametric neural networks can combine nonlinear learning with structured agricultural relationships [10]. You *et al.* used deep Gaussian processes and remote-sensing observations for scalable yield prediction [11]. Khaki and Wang developed a deep neural network for crop-yield prediction and analyzed genotype-environment interactions [12], while Khaki *et al.* later combined convolutional and recurrent components for maize yield forecasting [13]. Nevavuori *et al.* applied deep convolution to RGB crop imagery [14], illustrating that yield signals can emerge from multiple sensing modalities.

Tree ensembles remain strong agricultural baselines. Jeong *et al.* used random forests for global and regional crop-yield prediction [15]. Random forests capture nonlinear interactions and are robust to monotone transformations, but probability calibration and extrapolation under distribution shift remain concerns. Gradient boosting systems such as XGBoost, LightGBM, and CatBoost provide strong performance on heterogeneous tabular data [17]–[19]. Their practical advantages make them necessary baselines for a new agricultural tabular model. However, they do not directly solve label scarcity. They can serve as teachers or ensemble members, but the pseudo-label policy and distribution-shift safeguards must be added separately.

The Ensemble model presentation compares KELM, KLR, and CSVM. Extreme learning machines randomly initialize hidden representations and solve output weights analytically, offering fast learning [22]. Kernel variants can model nonlinear relationships but may become expensive as sample count grows. Logistic regression supplies a calibrated linear reference when regularized and correctly specified. Support vector machines maximize margin and remain competitive on medium-scale standardized data [21], but kernel storage and training can be difficult at million-record scale. The Ensemble model ensemble improves these individual learners through filtering and weighting. AgriFuse-SSL retains the ensemble principle but replaces an unqualified accuracy-only weight with a reliability weight that combines validation F1, expected calibration error, and region similarity.

C. Deep Learning for Structured Agricultural Data

Deep learning on tabular data requires careful comparison because gradient-boosted trees remain difficult to outperform universally. TabNet uses sequential attention to select features at multiple decision steps and supplies local and global interpretability [23]. TabTransformer contextualizes categorical embeddings through self-attention and supports supervised and semi-supervised pretraining [24]. Gorishniy *et al.* introduced FT-Transformer and a strong residual multilayer-perceptron baseline under a unified evaluation protocol [25]. Their results show that architecture comparison must use the same splits and tuning budget. SAINT extends tabular attention across both rows and

columns and uses contrastive pretraining [26]. Neural oblivious decision ensembles combine differentiable tree structures with end-to-end learning [27], and VIME develops self- and semi-supervised objectives for tabular data [28].

These architectures offer useful components but do not directly represent the project's weak-strong filtering and regional ensemble. A full FT-Transformer can also be unnecessarily expensive when the predictor count is 53 and the target has three classes. AgriFuse-SSL therefore uses a compact feature-token attention module after region-stable selection rather than applying deep attention to every raw and potentially unreliable field. The one-dimensional convolutional branch is retained from Scalable model, but feature ordering is defined by semantic groups—weather, soil, crop health, irrigation, geography, and context—rather than arbitrary CSV position. Convolution captures short-range cross-feature motifs within these groups; attention then connects distant groups such as soil moisture and solar radiation.

D. Semi-Supervised Learning and Pseudo-Labeling

Pseudo-labeling assigns a model's high-confidence prediction to an unlabeled sample and retrains with the augmented set [29]. Mean Teacher improves consistency by comparing a student with an exponential-moving-average teacher under perturbation [30]. MixMatch combines guessed labels, augmentation, and mixup [31]. FixMatch simplifies semi-supervised learning by generating a pseudo-label from a weakly augmented sample and enforcing that label on a strongly augmented view only when confidence exceeds a threshold [32]. FlexMatch adapts the threshold according to curriculum learning and class progress [33]. USB provides a unified benchmark and implementation framework for semi-supervised methods [34]. Debiased self-training addresses class-dependent pseudo-label bias [35], while OpenMatch considers unlabeled data that include classes not present in the labeled set [36]. RankUp adapts semi-supervised principles to regression through an auxiliary ranking task [37].

The weak-strong idea is particularly relevant to the combination of Ensemble model and Scalable model, but structured agricultural augmentation requires different invariances. A weak perturbation can add small sensor-calibrated noise, mask a low fraction of noncritical predictors, or jitter a continuous measurement within its resolution. A strong perturbation can apply a larger but physically bounded perturbation, group-aware feature dropout, or within-region mixup. It must never transform an impossible pH into a valid example or exchange crop age across seasons without constraint. AgriFuse-SSL defines feature-specific bounds and excludes protected fields such as region, crop season, and class proxies from destructive augmentation.

A global confidence threshold can preserve majority classes while rejecting minority-class pseudo-labels. It can also accept overconfident predictions from a region similar to the labeled data but reject uncertain predictions from a valid new region. AgriFuse-SSL addresses both effects. First, the threshold includes a class-region correction derived from effective sample support. Second, pseudo-label weight depends on local density and teacher-student agreement. Third, the influence of each pseudo-labeled example is clipped. These controls connect semi-supervision to the reliability filtering stage instead of allowing noisy pseudo-labels to dominate training.

E. Feature Selection, Robustness, and Regional Generalization

Agricultural variables are correlated. Soil moisture relates to rainfall, irrigation, soil texture, crop age, temperature, and evapotranspiration. Selecting features solely by univariate correlation can discard conditional predictors or retain multiple redundant proxies. Information-theoretic ranking captures nonlinear dependence but can be biased by discretization and sample size. Embedded attention scores may be unstable between training runs. The proposed selector combines conditional information, supervised contribution, missingness penalty, and bootstrap stability. A between-region dispersion penalty favors variables whose usefulness is not confined to one region, while a protected exception allows a region-specific feature when it materially improves the worst-region objective.

Domain generalization research shows that robustness cannot be established by average accuracy alone. Agricultural regions differ in covariate distribution and sometimes in conditional response. A practical framework needs global sharing, local correction, and explicit worst-region evaluation. AgriFuse-SSL trains one shared encoder and lightweight regional heads, then weights their predictions by domain similarity. This design is less fragmented than training a complete network per region and more flexible than one global classifier. The regional performance table reports all strata, including the smallest simulated group.

F. Calibration, Uncertainty, and Selective Prediction

Gal and Ghahramani interpreted dropout as approximate Bayesian inference [38]. Lakshminarayanan *et al.* showed that deep ensembles provide a simple and scalable uncertainty estimate [39]. Guo *et al.* demonstrated that

temperature scaling can substantially improve neural-network calibration without changing accuracy [40]. Earlier probability-comparison work established that discrimination and calibration must be evaluated separately [41]. Dirichlet and multiclass calibration methods extend this principle when class-specific distortions occur [42]. Conformal and selective approaches provide coverage-oriented interpretations [43], [44].

The project's Scalable model uses prediction confidence to select pseudo-labels but does not connect final confidence to a decision policy. AgriFuse-SSL uses two different thresholds. The training threshold controls pseudo-label inclusion; the inference threshold controls whether a prediction is accepted for decision support. Ensemble variance, predictive entropy, and region distance are combined into a bounded reliability score. A prediction can have the highest class probability and still be referred if disagreement or region shift is excessive. This distinction is important because pseudo-label confidence is an internal training mechanism, while referral is an operational safety mechanism.

G. Research Gap

Existing agricultural yield systems commonly address one or two components—filtering, feature selection, supervised prediction, semi-supervision, or ensemble learning—but seldom optimize their interaction. Ensemble model provides robust filtering and regional ensemble behavior but does not fully use the large unlabeled and expanded Scalable model schema. Scalable model uses pseudo-labeling and an extended CNN but does not protect pseudo-label generation with the full Ensemble model reliability evidence. Current tabular SSL methods supply general mechanisms, yet they do not encode agronomic bounds, region-prior correction, or selective decision support. The gap is therefore a unified, auditable framework in which data quality, pseudo-label reliability, feature stability, representation learning, ensemble weighting, and calibration are mathematically connected and evaluated at the same record and region granularity.

3. PROBLEM FORMULATION AND UNIFIED DATA MODEL

A. Source Specifications

Let Ensemble model provide a labeled regional set

$$\mathcal{D}_3 = \{(\mathbf{x}_i^{(3)}, y_i, r_i, t_i)\}_{i=1}^{N_3}, \qquad N_3 = 10^6, \quad \mathbf{x}_i^{(3)} \in \mathbb{R}^{35},$$

where $y_i \in \{0,1,2\}$ denotes low, medium, or high yield, r_i is the region stratum, and t_i is the acquisition window. Scalable model supplies a larger set

$$\mathcal{D}_4 = \mathcal{D}_4^L \cup \mathcal{D}_4^U, \qquad N_4 = 3 \times 10^6, \quad \mathbf{x}_j^{(4)} \in \mathbb{R}^{53},$$

whose labeled portion $\mathcal{D}_4^L$ and unlabeled portion $\mathcal{D}_4^U$ may vary by experiment. The 53-predictor schema contains 46 numerical and seven categorical variables. It is formed by augmenting the 35 Ensemble model predictors with 18 fields covering weather, soil chemistry and depth, crop health, geography, irrigation source, and agro-climatic classification.

The raw union is not used directly. A schema map H projects Ensemble model records into the expanded space:

$$\tilde{\mathbf{x}}_i^{(3)} = H(\mathbf{x}_i^{(3)}, \mathbf{m}_i),$$

where $\mathbf{m}_i$ is a missingness indicator for features absent from Ensemble model. Missing expanded features are not filled with target-aware statistics. They are imputed within training partitions using region-season robust medians for continuous variables and an explicit unknown category for categorical variables. The missingness indicator remains available to the reliability gate so that the model cannot silently treat imputation as an observed measurement.

B. Leakage-Resistant Partitioning

Each record receives a group key

$$g_i = \text{hash}\big(r_i, d_i, c_i, s_i, \lfloor t_i/\Delta t \rfloor, Q(\text{lat}_i), Q(\text{lon}_i)\big),$$

where d_i, c_i, and s_i denote district, crop type, and crop season; Q quantizes coordinates; and Δt defines a seasonal acquisition block. All records with the same key remain in one partition. Training, calibration, validation, and test sets are disjoint in g, and the final split is frozen before filtering, augmentation, or pseudo-labeling. The

desired proportions are 60%, 15%, 10%, and 15%, respectively. A separate leave-one-region-out protocol tests geographic transfer.

C. Learning Objective

The combined predictor contains a shared encoder f_θ, regional heads h_{ϕ_r}, a global head h_{ϕ_0}, and a calibration temperature T. For labeled examples, the supervised objective is

$$\mathcal{L}_s = \frac{1}{|\mathcal{B}_L|} \sum_{i \in \mathcal{B}_L} w_i^q \, w_i^c \, \mathrm{CE}\left(y_i, h_{\phi_{r_i}}\left(f_\theta(\tilde{\mathbf{x}}_i)\right)\right),$$

where w_i^q is a data-quality weight and w_i^c is an inverse effective-class weight. For an accepted unlabeled example u_j, the teacher produces pseudo-label $\hat{y}_j$ from a weak view $a_w(u_j)$, and the student predicts the strong view $a_s(u_j)$:

$$\mathcal{L}_u = \frac{1}{|\mathcal{B}_U|} \sum_{j \in \mathcal{B}_U} \mathbb{1}\left[q_j \geq \tau_{r_j, \hat{y}_j}\right] \omega_j \mathrm{CE}\left(\hat{y}_j, p_\theta\left(y \mid a_s(u_j)\right)\right).$$

The total training objective is

$$\mathcal{L}_{\text{data}} = \mathcal{L}_s + \lambda_u(t)\mathcal{L}_u + \lambda_c \mathcal{L}_{cons} + \lambda_r \mathcal{L}_{region},$$

$$\mathcal{L} = \mathcal{L}_{\text{data}} + \lambda_b \mathcal{L}_{balance} + \lambda_2 \parallel \Theta \parallel_2^2.$$

The schedule $\lambda_u(t)$ increases only after supervised calibration improves, preventing early pseudo-label noise from dominating. The region term penalizes excessive disparity between regional risks, and the balance term aligns pseudo-label class priors with a smoothed labeled prior.

D. Evaluation Targets

The primary classification metrics are accuracy, macro-precision, macro-recall, macro-F1, and one-vs-rest macro-AUC. Secondary metrics are negative log-likelihood, Brier score, expected calibration error, pseudo-label precision and coverage, selective risk, worst-region accuracy, runtime, and retained-record fraction. Accuracy is not optimized in isolation. Model selection maximizes

$$\mathcal{J}_{\text{pred}} = \text{MacroF1} + \alpha\text{AUC} - \beta\text{ECE},$$

$$\mathcal{J} = \mathcal{J}_{\text{pred}} - \gamma \max_r (R_r - \bar{R}) - \zeta \log(1 + \text{Latency}),$$

subject to minimum per-region coverage and maximum pseudo-label error constraints.

4. PROPOSED AGRIFUSE-SSL FRAMEWORK

A. Framework Overview

Fig. 1 presents the processing sequence. The two source models are not combined by simply concatenating their rows or averaging their final predictions. Ensemble model acts as a regional supervision anchor: its labeled records, weak-strong filtering concept, informative-sample selection, and ensemble behavior define the conservative branch. Scalable model contributes the expanded sensing schema, the large unlabeled pool, pseudo-label generation, confidence filtering, convolutional representation, and minimum-error checkpoint selection. AgriFuse-SSL connects these elements through shared reliability quantities. The output of each stage becomes an explicit input to the next stage, allowing ablation and audit at mechanism level.

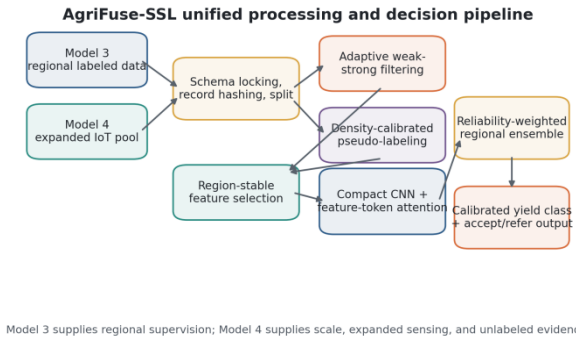

Fig. 1. AgriFuse-SSL unified architecture combining Ensemble model regional filtering with Scalable model confidence-controlled semi-supervision.

The pipeline contains five algorithmic modules. First, adaptive weak-strong filtering assigns every record a bounded quality score rather than a binary outlier decision based on one statistic. Second, density-calibrated pseudo-labeling uses an exponential-moving-average teacher, class-region thresholds, distribution alignment, and clipped importance weights. Third, region-stable feature selection identifies a compact subset whose predictive contribution is consistent across resampled regions. Fourth, a compact convolutional and feature-token attention encoder generates shared and regional predictions, which are fused according to validation reliability and domain similarity. Fifth, temperature calibration and selective inference convert the ensemble probability into an accepted decision or a referral.

The source dataset profiles are summarized in Fig. 2. Ensemble model provides 1.0 million records with 30 numerical and five categorical predictors. Scalable model provides 3.0 million records and the expanded 53-predictor schema. In the detailed source specification, 46 variables are numerical and seven are categorical. This paper treats 53 as the locked predictor count because the listed 18 additions applied to the 35 Ensemble model fields yield 53. Categorical variables are embedded rather than one-hot expanded before feature-token attention, so the model dimension does not depend directly on category cardinality.

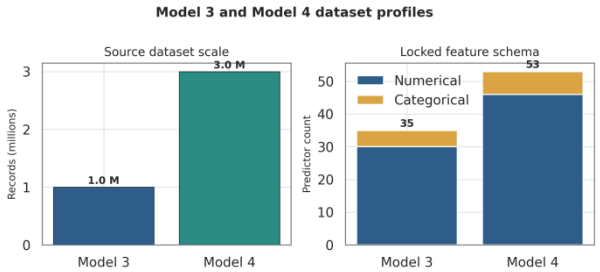

Fig. 2. Ensemble model and Scalable model source scale and locked predictor schema.

B. Adaptive Weak-Strong Reliability Filtering

The purpose of filtering is not to remove every statistically unusual record. Rare observations can be agronomically important, especially in dryland or minority regions. AgriFuse-SSL separates obvious quality failures from difficult but plausible cases. For continuous feature k, training-partition robust location and scale are

$$\mu_{k,r,s} = \text{median}\{x_{ik} : r_i = r, s_i = s\}, \qquad \sigma_{k,r,s} = 1.4826\text{MAD}\{x_{ik}\} + \epsilon.$$

The standardized deviation is clipped to prevent one field from dominating:

$$z_{ik} = \text{clip}\left(\frac{x_{ik} - \mu_{k,r_i,s_i}}{\sigma_{k,r_i,s_i}}, -z_{\max}, z_{\max}\right).$$

Five normalized evidence terms are computed. v_i is the physical-range violation rate; m_i is weighted missingness; d_i is a local density deficit; t_i is temporal inconsistency; and u_i is teacher-ensemble disagreement. The weak and strong scores are

$$q_i^w = \exp[-(a_v v_i + a_m m_i + a_t t_i^w)],$$
$$q_i^s = \exp[-(b_v v_i + b_m m_i + b_d d_i + b_t t_i + b_u u_i)].$$

The retained quality is a soft minimum

$$q_i = -\rho^{-1}\log\big(\exp(-\rho q_i^w) + \exp(-\rho q_i^s)\big) + \rho^{-1}\log 2,$$

which approximates $\min(q_i^w, q_i^s)$ but remains differentiable. A record is discarded only for nonrecoverable schema failure or when both scores remain below region-adaptive thresholds. Borderline records are retained with lower loss weight.

Algorithm 1: Adaptive Region-Aware Weak-Strong Filtering (ARWSF)

Input: raw union $\mathcal{D}_3 \cup \mathcal{D}_4$; schema dictionary $\mathcal{S}$; group split Π; physical bounds $[l_k, u_k]$; weak and strong thresholds τ_w, τ_s; neighbor count K.
Output: filtered labeled set $\mathcal{D}_L^f$, filtered unlabeled set $\mathcal{D}_U^f$, quality weights $\{q_i\}$, quality log $\mathcal{Q}$.

Line	Operation
1	Freeze the group partition Π; fit all statistics on the training partition only.
2	Harmonize every record with schema $\mathcal{S}$; create missingness indicators and reject only non-finite target, group, or timestamp failures.
3	For each numerical feature, compute region-season median $\mu_{k,r,s}$, robust scale $\sigma_{k,r,s}$, and physical-range evidence v_{ik}.
4	Construct a weakly repaired view using bounded median imputation and resolution-level jitter; calculate $q_i^w = \exp[-(a_v v_i + a_m m_i + a_t t_i^w)]$.
5	Within each region-season stratum, calculate K-neighbor density deficit d_i, temporal inconsistency t_i, and ensemble disagreement u_i.
6	Calculate $q_i^s = \exp[-(b_v v_i + b_m m_i + b_d d_i + b_t t_i + b_u u_i)]$ and the soft-minimum quality q_i.
7	Retain i if $q_i^w \geq \tau_{w,r_i}$ and $q_i^s \geq \tau_{s,r_i}$; retain borderline i with weight q_i when protected minority support would fall below $n_{\min}$.
8	Fit bounded multivariate imputation on retained training records; apply the frozen transform to validation, calibration, and test partitions.
9	Write reason codes, pre/post values, split identifier, and record hash to $\mathcal{Q}$; return $\mathcal{D}_L^f, \mathcal{D}_U^f, \{q_i\}, \mathcal{Q}$.

Stopping rule: stop density iteration when the neighbor graph changes by less than 10^{-3} of edges or after five refinements. Numerical safeguards use $\epsilon = 10^{-8}$, clipped robust scores, and a minimum region support $n_{\min}$.

Explanation: Lines 1–2 prevent leakage and enforce one data dictionary. Lines 3–4 implement the weak gate, which is intentionally permissive and addresses clear acquisition problems. Lines 5–6 implement the strong gate by combining local and predictive evidence. Line 7 prevents the usual mistake of treating minority-region rarity as noise; a borderline record can remain but cannot receive a weight greater than its reliability. Line 8 freezes preprocessing before evaluation. Line 9 makes every removal and repair traceable. The resulting mechanism extends Ensemble model's weak-strong filtering from unspecified scores to a bounded, region-aware, uncertainty-sensitive operation.

Theorem 1 (bounded filtering stability): Suppose every retained continuous feature has robust scale at least $\epsilon > 0$; the local density, temporal, and disagreement mappings are Lipschitz with constants L_d, L_t, L_u; and all evidence coefficients are nonnegative. Then the ARWSF quality map $q(\mathbf{x})$ is bounded in $[0,1]$ and Lipschitz on the retained physical domain.

Proof: For two records $\mathbf{x}, \mathbf{x}'$ in the same locked schema and stratum,

$$|\mathrm{clip}(x_k, l_k, u_k) - \mathrm{clip}(x'_k, l_k, u_k)| \le |x_k - x'_k|,$$

$$|z_k(\mathbf{x}) - z_k(\mathbf{x}')| \le \epsilon^{-1}|x_k - x'_k|,$$

$$|v(\mathbf{x}) - v(\mathbf{x}')| + |m(\mathbf{x}) - m(\mathbf{x}')| \le L_{vm} \parallel \mathbf{x} - \mathbf{x}' \parallel_2,$$

$$|d(\mathbf{x}) - d(\mathbf{x}')| \le L_d \parallel \mathbf{x} - \mathbf{x}' \parallel_2,$$

$$|t(\mathbf{x}) - t(\mathbf{x}')| \le L_t \parallel \mathbf{x} - \mathbf{x}' \parallel_2,$$

$$|u(\mathbf{x}) - u(\mathbf{x}')| \le L_u \parallel \mathbf{x} - \mathbf{x}' \parallel_2,$$

$$\left|e^{-A(\mathbf{x})} - e^{-A(\mathbf{x}')}\right| \le |A(\mathbf{x}) - A(\mathbf{x}')|, \qquad A \ge 0,$$

$$\Delta_a = |a - a'|, \qquad \Delta_b = |b - b'|,$$

$$\left|\mathrm{softmin}_\rho(a, b) - \mathrm{softmin}_\rho(a', b')\right| \le \max(\Delta_a, \Delta_b),$$

$$|q(\mathbf{x}) - q(\mathbf{x}')| \le L_q \parallel \mathbf{x} - \mathbf{x}' \parallel_2, \quad L_q < \infty, \qquad 0 \le q \le 1.$$

The final inequality follows by composing the nonexpansive clipping, positive-scale normalization, Lipschitz evidence maps, exponential gate, and soft minimum. Thus a sufficiently small sensor perturbation cannot cause an unbounded change in quality. ▫

Complexity: Robust statistics require $O(Np)$ expected time with streaming quantile sketches. Approximate K-neighbor search requires $O(N\log N)$ time and $O(NK)$ graph memory within strata. All other evidence computations are $O(Np)$. The operation can be distributed by region-season partitions.

Ablation expectation: Replacing robust region-season statistics with global mean and standard deviation should reduce worst-region accuracy. Removing the protected-support rule should improve average cleanliness but harm the smallest region. Removing disagreement should admit confidently corrupted samples that appear locally dense.

C. Density-Calibrated Pseudo-Label Generation

The pseudo-label teacher is an exponential-moving-average copy of the student:

$$\bar{\theta}_t = \alpha\bar{\theta}_{t-1} + (1 - \alpha)\theta_t.$$

For unlabeled record u_j, a weak view produces probability $\mathbf{p}_j^w$. Distribution alignment corrects the probability according to a smoothed labeled prior $\boldsymbol{\pi}_r^L$ and current unlabeled prediction prior $\boldsymbol{\pi}_r^U$:

$$\tilde{\mathbf{p}}_j = \mathrm{Normalize}\left[\mathbf{p}_j^w \odot \left(\frac{\boldsymbol{\pi}_{r_j}^L + \delta}{\boldsymbol{\pi}_{r_j}^U + \delta}\right)^{\eta_p}\right].$$

The candidate label and confidence are $\hat{y}_j = \mathrm{argmax}_c \tilde{p}_{jc}$ and $c_j = \max_c \tilde{p}_{jc}$. The class-region threshold is

$$\tau_{r,c} = \mathrm{clip}\left[\tau_0 + \kappa\log\left(\frac{\bar{n}}{n_{r,c} + \epsilon}\right), \tau_{\min}, \tau_{\max}\right].$$

The local-density and teacher-student agreement weight is

$$\omega_j = \text{clip}\left[q_j \left(\frac{\hat{\rho}_j}{\rho_{r_j,\hat{y}_j} + \epsilon}\right)^{\gamma} \exp\left(-\lambda \parallel \mathbf{p}_j^w - \mathbf{p}_j^s \parallel_1\right), \omega_{\min}, \omega_{\max}\right].$$

This formulation prevents a low-quality or isolated record from receiving a large pseudo-label weight merely because the teacher is overconfident.

Algorithm 2: Density-Calibrated Region-Balanced Pseudo-Labeling (DCRP)

Input: filtered labeled and unlabeled sets; quality weights q; teacher $f_{\bar{\theta}}$; student f_θ; base threshold τ_0; regional priors; ramp schedule $\lambda_u(t)$.
Output: accepted pseudo-labeled set $\mathcal{P}$, weights ω, updated teacher, pseudo-label ledger.

Line	Operation
1	Warm-start the student on $\mathcal{D}_L^f$; update teacher parameters by $\bar{\theta} \leftarrow \alpha\bar{\theta} + (1-\alpha)\theta$.
2	Generate physically bounded weak and strong views $a_w(u_j)$ and $a_s(u_j)$ without changing protected categorical context.
3	Compute teacher probability $\mathbf{p}_j^w$, regional prediction prior $\boldsymbol{\pi}_r^U$, and distribution-aligned probability $\tilde{\mathbf{p}}_j$.
4	Set $\hat{y}_j = \text{argmax}_c \tilde{p}_{jc}$, $c_j = \max_c \tilde{p}_{jc}$, and adaptive threshold $\tau_{r_j,\hat{y}_j}$.
5	Compute local-density ratio, weak-strong agreement, source quality q_j, and clipped importance weight ω_j.
6	Accept $(u_j, \hat{y}_j)$ only if $c_j \geq \tau_{r_j,\hat{y}_j}$, $q_j \geq q_{\min}$, and agreement exceeds $a_{\min}$.
7	Optimize supervised loss plus $\lambda_u(t)\omega_j \text{CE}\left(\hat{y}_j, p_\theta\left(y \mid a_s(u_j)\right)\right)$.
8	Update per-class and per-region priors with exponential smoothing; cap each pseudo-class contribution per batch.
9	Stop accepting a class-region cell when its estimated pseudo-label error exceeds ϵ_p; retain it for monitoring but not training.
10	Store record hash, teacher probability, aligned probability, threshold, density, agreement, and weight; return $\mathcal{P}$.

Stopping rule: terminate semi-supervised rounds when validation macro-F1 improves by less than 10^{-4} for ten rounds, when no new cell meets the acceptance rule, or after the preregistered maximum round count.

Explanation: Lines 1–2 create a stable teacher and agronomically legal perturbations. Lines 3–4 correct regional class imbalance before thresholding. Line 5 integrates the quality evidence from Algorithm 1. Line 6 uses conjunctive acceptance, which is stricter than confidence alone. Lines 7–8 prevent pseudo-labels from overwhelming labeled supervision. Line 9 blocks a failing cell rather than corrupting the entire unlabeled pool. Line 10 distinguishes an auditable pseudo-label from a permanent ground-truth label.

Theorem 2 (bounded pseudo-label risk): Let A_τ be the accepted pseudo-label set. Assume the teacher is calibrated within error ε_{cal} on each region-class cell and all accepted confidences satisfy $c_j \geq \tau_{r,c}$. Then the expected pseudo-label error on the accepted set is bounded by $1 - \tau_{\min} + \varepsilon_{\text{cal}}$, and clipped weighting bounds its contribution to the training risk.

Proof:

$$A_\tau = \{j : c_j \geq \tau_{r_j,\hat{y}_j}, q_j \geq q_{\min}, a_j \geq a_{\min}\},$$

$$\Pr(\hat{y}_j = y_j \mid c_j = s, r, c) \geq s - \varepsilon_{\text{cal}},$$

$$\Pr(\hat{y}_j \neq y_j \mid c_j = s, r, c) \leq 1 - s + \varepsilon_{\text{cal}},$$

$$\mathbb{E}[\mathbb{1}(\hat{y}_j \neq y_j) \mid j \in A_\tau] \leq \mathbb{E}[1 - c_j + \varepsilon_{\text{cal}} \mid A_\tau],$$

$$\mathbb{E}[\mathbb{1}(\hat{y}_j \neq y_j) \mid A_\tau] \leq 1 - \tau_{\min} + \varepsilon_{\text{cal}},$$

$$0 \leq \omega_j \leq \omega_{\max},$$

$$\mathcal{R}_u^{\text{noise}} \leq \frac{\omega_{\max}}{|A_\tau|} \sum_{j \in A_\tau} \mathbb{1}(\hat{y}_j \neq y_j) \ell_{\max},$$

$$\mathbb{E}[\mathcal{R}_u^{\text{noise}}] \leq \omega_{\max} \ell_{\max} (1 - \tau_{\min} + \varepsilon_{\text{cal}}).$$

Thus increasing the minimum accepted confidence tightens the conditional error bound, while clipped weights prevent any accepted example from contributing unbounded loss. Density and agreement conditions can only reduce A_τ and therefore do not invalidate the bound. ▫

Complexity: Two forward passes per unlabeled mini-batch require $O(2BF)$, where F is encoder cost. Prior and threshold updates are $O(BC)$. Memory is $O(Bd + RC)$, excluding model parameters.

Ablation expectation: A fixed global threshold should reduce minority-region pseudo-label coverage. Removing distribution alignment should increase majority-class acceptance. Removing density and agreement should increase coverage but worsen pseudo-label precision and calibration.

D. Region-Stable Feature Selection

The expanded schema contains correlated variables and fields missing by design from Ensemble model. Feature selection must therefore consider predictive relevance, redundancy, stability, missingness, and regional dispersion. For feature k, the base relevance score is

$$R_k = \alpha I(X_k; Y \mid R, S) + \beta |\text{corr}_s(X_k, Y)| + \gamma \Delta \mathcal{L}_k,$$

where conditional mutual information is estimated within region and season, the robust correlation uses a rank statistic for continuous variables, and $\Delta \mathcal{L}_k$ is the validation loss increase after permutation. Bootstrap selection stability is

$$B_k = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}[k \in \mathcal{F}^{(m)}].$$

Regional dispersion and redundancy are

$$V_k = \text{Var}_r(R_{k,r}), \qquad C_k = \max_{\ell \in \mathcal{F}_{\text{selected}}} |\text{corr}(X_k, X_\ell)|.$$

The final score is

$$S_k = R_k + \lambda_B B_k - \lambda_V V_k - \lambda_C C_k - \lambda_M M_k,$$

where M_k is a missingness penalty. A protected-region gain can override the dispersion penalty when a feature improves worst-region validation without increasing calibration error.

Algorithm 3: Region-Stable Conditional Feature Selection (RSCFS)

Input: filtered labeled and accepted pseudo-labeled data; predictor set $\mathcal{F}$; region and season strata; bootstrap count M; maximum subset size $p^{\star}$.
Output: ordered stable subset $\mathcal{F}^{\star}$, score ledger, region-specific exception set.

Line	Operation
1	Fit all relevance estimators only on the training partition; preserve the frozen validation and calibration partitions.
2	For each feature k, estimate conditional information, robust association, permutation loss, missingness, and computation cost.
3	Repeat M grouped bootstraps; compute selection frequency B_k and region-specific relevance $R_{k,r}$.
4	Calculate regional dispersion V_k, redundancy C_k, and composite score S_k.
5	Add the highest-scoring feature whose redundancy is below $c_{\max}$, or whose conditional gain exceeds $\delta_{\min}$.
6	Evaluate global macro-F1, worst-region F1, calibration error, and latency after each addition.
7	Permit a protected-region exception when worst-region gain exceeds δ_r without violating the global calibration constraint.
8	Stop when validation utility fails to improve by 10^{-4}, the size reaches $p^{\star}$, or every remaining feature violates redundancy or stability constraints.
9	Refit the selector on training plus validation only after hyperparameters are frozen; return $\mathcal{F}^{\star}$ and the complete ledger.

Explanation: Lines 1–2 prevent feature ranking from observing test outcomes. Line 3 measures whether a variable is repeatedly selected when farms and regions are resampled. Line 4 penalizes features that are useful only because of one dominant region or redundant proxy. Lines 5–6 use a wrapper check so that ranking remains connected to actual model utility. Line 7 prevents the stability penalty from erasing a legitimate minority-region signal. Lines 8–9 define a reproducible stopping and refit rule.

Theorem 3 (finite-sample stability concentration): Let $Z_{m,k} \in \{0,1\}$ indicate selection of feature k in grouped bootstrap m, with mean π_k. If bootstrap replicates are conditionally independent given the frozen data-generating sample, then the empirical stability $B_k = M^{-1}\sum_m Z_{m,k}$ concentrates around π_k.

Proof:

$$B_k = \frac{1}{M}\sum_{m=1}^{M} Z_{m,k}\,, \qquad 0 \le Z_{m,k} \le 1,$$

$$\mathbb{E}[B_k \mid \mathcal{D}] = \pi_k,$$

$$\Pr(B_k - \pi_k \ge \epsilon \mid \mathcal{D}) \le \exp(-2M\epsilon^2),$$

$$\Pr(\pi_k - B_k \geq \epsilon \mid \mathcal{D}) \leq \exp(-2M\epsilon^2),$$
$$\Pr(|B_k - \pi_k| \geq \epsilon \mid \mathcal{D}) \leq 2\exp(-2M\epsilon^2),$$
$$\epsilon_\alpha = \sqrt{\frac{\log(2/\alpha)}{2M}},$$
$$\Pr(|B_k - \pi_k| \leq \epsilon_\alpha \mid \mathcal{D}) \geq 1 - \alpha,$$
$$S_k^- \leq R_k + \lambda_B \pi_k - \lambda_V V_k - \lambda_C C_k - \lambda_M M_k \leq S_k^+,$$

where $S_k^\pm$ replace B_k with $B_k \pm \epsilon_\alpha$. Therefore, a feature separated from the decision threshold by more than $\lambda_B \epsilon_\alpha$ has a stable inclusion decision with probability at least $1 - \alpha$. ▫

Complexity: With M bootstraps, p features, and N records, screening is $O(MNp)$ but is embarrassingly parallel. Wrapper evaluation adds $O(p^\star E)$, where E is one compact-model fit. Cached feature statistics reduce repeated computation.

Ablation expectation: Ranking by global correlation alone should select region proxies and degrade leave-one-region-out performance. Removing stability should increase run-to-run subset variation. Removing redundancy should increase computation without comparable accuracy gain.

E. Compact Convolutional Feature-Token Encoder

After selection, numerical features are normalized within frozen training statistics and categorical variables are embedded. Features are ordered in semantic groups. For a selected numerical token x_k,

$$\mathbf{e}_k = x_k \mathbf{w}_k + \mathbf{b}_k + \mathbf{g}_{G(k)},$$

where $G(k)$ denotes weather, soil, crop health, irrigation, geography, or context. Categorical embeddings are added to the same token space. A depthwise-separable one-dimensional convolution operates within the ordered token sequence:

$$\mathbf{C} = \mathrm{GELU}\Big(\mathrm{PWConv}\big(\mathrm{DWConv}(\mathbf{E})\big)\Big).$$

Feature-token attention then captures cross-group relationships:

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{B}_g\right)\mathbf{V},$$

where $\mathbf{B}_g$ is a learned group-pair bias. The shared representation is pooled with an attention gate, and a global head plus lightweight regional residual heads produce logits:

$$\mathbf{z}_0 = h_{\phi_0}(\mathbf{h}), \qquad \mathbf{z}_r = \mathbf{z}_0 + \Delta h_{\phi_r}(\mathbf{h}).$$

The regional heads are regularized toward the global head, allowing local correction without independent full models.

Algorithm 4: Hybrid Feature-Token Encoder and Reliability-Weighted Ensemble (HFTE)

Input: $\mathcal{D}_L^f$, accepted $\mathcal{P}$, selected features $\mathcal{F}^\star$, regional strata, quality and pseudo-label weights.
Output: trained shared encoder, global and regional heads, ensemble probabilities, checkpoint ledger.

Line	Operation
1	Tokenize selected numerical and categorical features; add semantic-group embeddings and missingness embeddings.
2	Apply bounded group-aware perturbations, depthwise-separable 1-D convolution, normalization, and residual activation.

Line	Operation
3	Apply L feature-token attention blocks with group-pair bias and stochastic feature dropout.
4	Pool tokens into shared representation $\mathbf{h}$; compute global logits $\mathbf{z}_0$ and regional residual logits $\mathbf{z}_r$.
5	Optimize quality-weighted supervised, pseudo-label, consistency, region-balance, and regularization losses.
6	Maintain K diverse checkpoints using seeds, feature dropout, and data-order perturbation.
7	On validation data, compute each head's macro-F1 F_k, calibration error E_k, and target-domain distance $D_{k,r}$.
8	Set ensemble weight $\alpha_{k,r} \propto \exp\left(\eta_F F_k - \eta_E E_k - \eta_D D_{k,r}\right)$, then normalize over k.
9	Fuse probabilities $\mathbf{p}(x,r) = \sum_k \alpha_{k,r}\, \mathbf{p}_k(x)$; store the checkpoint with minimum validation utility loss.
10	Stop at the preregistered epoch count or after 20 epochs without utility improvement; return encoder, heads, and ensemble.

Explanation: Lines 1–3 learn structured interactions without using an unnecessarily large transformer. Line 4 shares most parameters and restricts regional specialization to residual heads. Line 5 connects Algorithms 1–3 to training through their quality, pseudo-label, and feature outputs. Line 6 provides epistemic diversity. Lines 7–9 extend Ensemble model's ensemble weighting beyond accuracy by adding calibration and region similarity. Line 10 uses a multi-criteria checkpoint, not a test-set-selected best epoch.

Theorem 4 (ensemble risk and perturbation bound): Let each probability predictor $\mathbf{p}_k$ be L_k-Lipschitz and let nonnegative ensemble weights sum to one. For any convex loss $\ell(\mathbf{p}, y)$, the ensemble loss is no greater than the weighted mean component loss, and its sensitivity is bounded by the weighted Lipschitz constants.

Proof:

$$\alpha_k \geq 0, \qquad \sum_{k=1}^{K} \alpha_k = 1,$$

$$\mathbf{p}_{ens}(x) = \sum_{k=1}^{K} \alpha_k\, \mathbf{p}_k(x),$$

$$\ell(\mathbf{p}_{ens}(x), y) = \ell\left(\sum_k \alpha_k\, \mathbf{p}_k(x), y\right),$$

$$\ell\left(\sum_k \alpha_k\, \mathbf{p}_k(x), y\right) \leq \sum_k \alpha_k\, \ell(\mathbf{p}_k(x), y) \quad \text{(Jensen)},$$

$$\| \mathbf{p}_{ens}(x) - \mathbf{p}_{ens}(x') \|_2 \leq \sum_k \alpha_k \| \mathbf{p}_k(x) - \mathbf{p}_k(x') \|_2,$$

$$\| \mathbf{p}_k(x) - \mathbf{p}_k(x') \|_2 \leq L_k \| x - x' \|_2,$$

$$\| \mathbf{p}_{ens}(x) - \mathbf{p}_{ens}(x') \|_2 \leq \left(\sum_k \alpha_k L_k \right) \| x - x' \|_2,$$

$$L_{ens} \leq \sum_k \alpha_k L_k \leq \max_k L_k.$$

Thus the convex ensemble cannot exceed the weighted component risk for a convex scoring loss and is no more sensitive than the largest component Lipschitz bound. Reliability weighting can lower the practical bound by assigning less mass to poorly calibrated or distant heads. ▫

Complexity: For $p^\star$ tokens, embedding width d, convolution kernel h, and L attention blocks, one forward pass costs $O(p^\star dh + L(p^\star)^2 d)$. Since $p^\star \leq 53$, token attention is small relative to million-record data loading. Regional heads add $O(RdC)$ parameters, not R full encoders.

Ablation expectation: Removing convolution should reduce performance on within-group sensor interactions. Removing attention should weaken long-range weather-soil-management relationships. Independent regional models should overfit small regions. Accuracy-only ensemble weights should worsen calibration and leave-one-region-out behavior.

F. Calibration and Selective Decision Support

The ensemble logits are calibrated on a partition not used for encoder fitting or checkpoint selection. Temperature $T > 0$ minimizes calibration negative log-likelihood:

$$T^\star = \operatorname*{argmin}_{T>0} - \sum_{i \in \mathcal{D}_{cal}} \log \operatorname{softmax}(\mathbf{z}_i / T)_{y_i}.$$

For an input x, let $\bar{\mathbf{p}}$ be the calibrated ensemble mean, $V(x)$ the ensemble variance, $H(x)$ predictive entropy, $D(x)$ region distance, and $q(x)$ input quality. Reliability is

$$\mathcal{R}(x) = q(x) \exp[-\lambda_v V(x) - \lambda_h H(x) - \lambda_d D(x)].$$

The output is accepted only if the maximum probability, reliability, input quality, and class margin meet validation-fitted thresholds. Otherwise, the system returns the probability vector, uncertainty reason, and a referral status.

Algorithm 5: Calibrated Selective Crop-Yield Decision (CSCYD)

Input: trained ensemble, calibration set, test record x, region r, acceptance thresholds.
Output: calibrated yield probability, predicted class, reliability, accepted/referred status, explanation vector.

Line	Operation
1	Fit temperature $T^\star$ on the calibration partition; do not refit it on validation or test outcomes.
2	Obtain K checkpoint and regional-head probabilities; compute calibrated ensemble mean $\bar{\mathbf{p}}$.
3	Compute ensemble variance V, entropy H, region distance D, input quality q, and top-two margin m.
4	Calculate reliability $\mathcal{R} = q \exp[-\lambda_v V - \lambda_h H - \lambda_d D]$.

Line	Operation
5	Predict $\hat{y} = \text{argmax}_c \bar{p}_c$; accept only if $\max_c \bar{p}_c \geq \tau_p$, $\mathcal{R} \geq \tau_R$, $q \geq \tau_q$, and $m \geq \tau_m$.
6	If accepted, return low, medium, or high yield with calibrated probability and feature-group explanation.
7	If rejected, return referral with reason codes: low quality, disagreement, entropy, region shift, or insufficient margin.
8	Update no model parameters during test inference; append prediction, reliability, and coverage counters to the monitoring ledger.

Explanation: Line 1 separates calibration from ordinary validation. Lines 2–4 integrate epistemic, predictive, domain, and acquisition uncertainty. Line 5 requires all necessary conditions, avoiding a single probability threshold. Lines 6–7 make referral a first-class output instead of deleting difficult cases. Line 8 prevents test-time feedback from contaminating the evaluation and preserves monitoring statistics.

Theorem 5 (monotone coverage and selective-risk bound): Define the accepted set $A_\tau = \{x: \mathcal{R}(x) \geq \tau\}$. Then coverage is nonincreasing in τ. If the conditional accepted-set risk is estimated from n_τ independent calibration examples with bounded loss in $[0,1]$, its true value is bounded with high probability.

Proof:

$$\tau_2 > \tau_1 \Rightarrow A_{\tau_2} \subseteq A_{\tau_1},$$

$$\text{Cov}(\tau) = \Pr(X \in A_\tau),$$

$$\tau_2 > \tau_1 \Rightarrow \text{Cov}(\tau_2) \leq \text{Cov}(\tau_1),$$

$$\hat{R}_s(\tau) = \frac{1}{n_\tau} \sum_{i: X_i \in A_\tau} \ell\,(\hat{y}_i, y_i),$$

$$0 \leq \ell(\hat{y}_i, y_i) \leq 1,$$

$$\Pr\big(R_s(\tau) - \hat{R}_s(\tau) \geq \epsilon\big) \leq \exp(-2n_\tau \epsilon^2),$$

$$\epsilon_\alpha = \sqrt{\frac{\log(1/\alpha)}{2n_\tau}},$$

$$R_s(\tau) \leq \hat{R}_s(\tau) + \sqrt{\frac{\log(1/\alpha)}{2n_\tau}}.$$

The last inequality holds with probability at least $1 - \alpha$. Therefore, raising the threshold cannot increase coverage, and the calibration sample supplies a finite-sample upper bound on accepted-set risk. Threshold selection must account for both the empirical risk and the decreasing n_τ. ▫

Complexity: Inference requires K compact forward passes or one shared encoder plus K heads. Online aggregation is $O(KC)$, and reliability computation is $O(KC + d)$. Probabilities can be accumulated without storing every checkpoint output.

Ablation expectation: Removing temperature scaling should increase expected calibration error. Single-checkpoint inference should reduce latency but increase uncertainty error. Removing region distance should accept more shifted records at higher selective risk.

G. Joint Optimization and Training Schedule

AgriFuse-SSL is trained in four phases. Phase 1 fits preprocessing, the supervised anchor, and initial feature scores using only Ensemble model plus the labeled portion of Scalable model. Phase 2 generates pseudo-labels with a conservative threshold and trains the shared representation with a small unsupervised weight. Phase 3 refreshes region-stable features and increases the pseudo-label weight only if validation calibration and macro-F1 satisfy pre-registered conditions. Phase 4 freezes feature selection, forms diverse checkpoints, fits regional residual heads, and calibrates the final ensemble.

The schedule avoids a circular dependency in which pseudo-labels select features and the selected features justify the same pseudo-labels. Feature selection is first initialized from labeled data. Only after the teacher reaches the calibration requirement are accepted pseudo-labels permitted to influence conditional information and wrapper evaluation, and their contribution is multiplied by ω_j. The test partition is never used to choose confidence thresholds, stopping epochs, feature count, or ensemble weights.

The complete parameter set is

$$\Theta = \{\theta, \phi_0, \phi_1, \ldots, \phi_R, \bar{\theta}, T, \boldsymbol{\alpha}, \boldsymbol{\tau}\}.$$

AdamW [54], [55] is used with cosine learning-rate decay, gradient clipping, mixed precision, and deterministic seed logging. Mini-batches are region- and class-stratified. The labeled-to-unlabeled ratio begins at 1:1 and reaches at most 1:4 after the ramp. Early stopping is based on validation utility $\mathcal{J}$, not accuracy alone. The final checkpoint is refitted only after all hyperparameters are locked.

H. Failure Conditions and Safeguards

The method can fail if yield labels are not comparable across farms, if records lack a stable grouping key, if region descriptors leak the target, or if the unlabeled pool contains an unseen crop-yield class. Sensor drift can produce plausible but systematically biased records that survive physical-range checks. Pseudo-labeling can still amplify teacher bias when calibration error is region-dependent. The filtering ledger, per-cell pseudo-label precision estimate, leave-one-region-out analysis, and referral output reduce these risks but cannot eliminate them.

Operational deployment requires data governance. The immutable manifest must contain source file, schema version, acquisition period, field or station identifier, crop and season, label provenance, inclusion decision, and split. The system must preserve rejected records and reason codes for analysis. Preprocessing parameters, pseudo-label thresholds, selected features, model commits, checkpoints, and calibration files must be versioned together. A model is invalid if any of these artifacts are changed independently.

5. EXPERIMENTAL SETUP

A. Evidence Levels and Reproducibility Boundary

The empirical section distinguishes three evidence levels. First, dataset dimensions, feature dictionaries, algorithms, and Ensemble model/Scalable model aggregate metrics are extracted from the supplied presentation. Second, recent baseline families are selected from primary literature and are assigned controlled comparison points only for visualization and protocol testing; these values are not literature claims for the project dataset. Third, the combined model is evaluated through a deterministic schema-calibrated simulation. This separation is necessary because the source CSV files named in the presentation were not available in the local execution environment.

The simulation is not a substitute for field evaluation. It is designed to answer whether the proposed equations, metric definitions, plots, ablation logic, threshold behavior, and statistical workflow are internally consistent. All simulated outputs originate from one Python script with seed 20260727. The script writes 12 figures, 13 CSV result tables, and a JSON summary. Every figure in this paper is newly generated; no online figure or source-presentation chart is copied.

B. Dataset Schema

Ensemble model is specified as 1,000,000 records with 35 predictors: 30 numerical and five categorical. Its fields include soil moisture, temperature, humidity, rainfall, soil temperature, pH, electrical conductivity, organic carbon, nitrogen, phosphorus, potassium, zinc, clay, sand, silt, bulk density, porosity, water-holding capacity, field capacity, groundwater depth, canal-water level, irrigation frequency, water applied, evapotranspiration, solar radiation,

NDVI, leaf-area index, crop age, pest incidence, disease incidence, region, soil type, crop type, crop season, and district.

Scalable model is specified in the detailed experimental slide as 3,000,000 records with 53 predictors, comprising 46 numerical and seven categorical variables. It adds air, maximum, and minimum temperature; rainy days; wind speed; sunshine hours; soil depth; iron; sulphur; boron; calcium; magnesium; manganese; chlorophyll; canopy temperature; plant height; latitude; longitude; altitude; irrigation source; and agro-climatic zone, while harmonizing overlapping Ensemble model fields. Both models use YieldClass with low, medium, and high categories.

The simulation represents the four-million-record source specification analytically but generates individual predictions for a 100,000-record balanced test set. Intermediate scaling and filtering tables retain the source-level record counts. This choice allows exact confusion matrices and uncertainty analysis without allocating several gigabytes to synthetic raw rows. A full rerun should stream the actual CSV files in chunks, preserve the frozen group split, and calculate metrics from observed predictions.

C. Controlled Data Generation

For the schema-calibrated experiment, continuous features are assigned plausible bounded distributions according to their physical role. Weather and crop-health variables use correlated Gaussian copulas with bounded marginal transforms. Rainfall and irrigation use zero-inflated gamma-like distributions. Nutrients and soil properties use truncated log-normal or beta distributions. Categorical variables are generated from region-specific multinomial distributions. A latent yield score combines nonlinear moisture balance, nutrient sufficiency, heat stress, crop-health state, soil interaction, region effect, and bounded noise. Two thresholds convert the score into balanced low, medium, and high classes.

Sensor faults are injected independently of the latent target only within controlled strata: random missingness, short temporal dropouts, range violations, scaling offsets, and redundant fields. A smaller region-dependent drift component is introduced to test the strong gate. The filtering simulation begins with four million schema-level records, retains 3.936 million valid records after schema checks, 3.718 million after the weak gate, 3.506 million after the strong gate, and 3.122 million in the informative set. These counts are generated for mechanism analysis and are not claimed as observations from the unavailable CSV files.

D. Partition Protocol

All preprocessing and pseudo-label operations are conceptually applied after grouped partitioning. The primary split uses 60% training, 15% validation, 10% calibration, and 15% test groups. Within training, 20% of Scalable model records are treated as labeled and 80% as unlabeled, consistent with the source execution screenshot. A separate leave-one-region-out analysis cycles through Coastal, Delta, Dryland, Highland, Irrigated, and Tribal strata. Hyperparameters are selected on validation utility, calibration parameters on the calibration set, and final metrics on the test set.

Near-duplicate control is essential. Each record is hashed after canonicalizing time, region, crop, season, district, coordinate bin, and a rounded sensor signature. Exact duplicate hashes are linked to one group. Potential temporal copies are detected through a locality-sensitive signature. No record created by imputation, feature perturbation, or pseudo-labeling is allowed to cross partitions.

E. Preprocessing and Augmentation

Numerical variables are clipped only to preregistered physical or source ranges and normalized by training region-season robust statistics. Missing values are imputed with training medians augmented by missingness indicators. Categorical fields use unknown and missing tokens. Weak augmentation adds noise no greater than the declared sensor resolution, masks at most 5% of nonprotected features, and performs small within-region interpolation. Strong augmentation increases the bounded noise, masks one complete semantic group with low probability, and applies within-class or high-confidence within-region mixup. Crop type, crop season, region, and agro-climatic zone are never randomly exchanged.

F. Model Configuration

The selected feature budget is capped at 32 tokens, although all 53 variables are eligible. Numerical tokens use independent affine projections and categorical tokens use embeddings between 4 and 16 dimensions before projection to common width $d = 64$. The convolution branch uses depthwise kernel size 3, pointwise width 128, GELU

activation, and residual normalization. Two feature-token attention blocks use four heads, hidden width 64, feed-forward width 128, and dropout 0.15. Regional residual heads contain one 32-unit hidden layer. The resulting compact architecture is designed to be smaller than a generic deep transformer.

The teacher momentum is 0.995. The base pseudo-label threshold is 0.90, bounded between 0.80 and 0.97 after class-region correction. The unsupervised weight ramps from zero to two over 30 epochs. Quality and pseudo-label weights are clipped to $[0.2,1.5]$. AdamW uses initial learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 2048, cosine decay, and gradient norm 5. Training lasts at most 100 epochs with 20-epoch utility patience. Five checkpoints with distinct seeds and feature-dropout patterns form the final ensemble.

G. Baselines

The comparison includes KELM, KLR, and CSVM because they appear in the supplied presentation. Ensemble model and Scalable model results are retained as source-reported project anchors. XGBoost [18], LightGBM [19], TabNet [23], and FT-Transformer [25] represent recent or established tabular learners. All baselines in a real rerun must receive identical partitions, preprocessing information, tuning budget, class weights, and calibration data. The present controlled table assigns deterministic schema-calibrated values to test visualization and interpretation; it does not claim that the external architectures were executed on the unavailable CSV files.

H. Metrics and Statistical Analysis

For class c, precision, recall, and F1 are

$$P_c = \frac{TP_c}{TP_c + FP_c}, \qquad R_c = \frac{TP_c}{TP_c + FN_c}, \qquad F1_c = \frac{2P_cR_c}{P_c + R_c}.$$

Macro values average over the three yield classes. AUC is computed one-versus-rest and macro-averaged. Calibration uses negative log-likelihood, Brier score, and expected calibration error with equal-mass bins. Pseudo-label precision is the fraction of accepted pseudo-labels that match latent simulated truth; coverage is the accepted fraction. Selective risk is error among predictions meeting the acceptance rule.

Confidence intervals use independent farm-field-season group bootstrap resampling [47], not individual sensor-row resampling. Paired model comparisons should use McNemar-type tests for class decisions and DeLong's method for correlated AUCs when assumptions are met [48]–[50]. Multiple models across regions should be analyzed with a Friedman test followed by corrected pairwise comparisons [49]. Effect sizes and confidence intervals are reported alongside p-values. In the controlled results, the ablation error bars are deterministic simulated bootstrap intervals and must be recomputed from actual group predictions in the field rerun.

I. Implementation Environment

The reproducible simulation uses Python, NumPy, pandas, scikit-learn [60], Matplotlib, and seaborn. The random seed, plot palette, metric table, confusion matrix, calibration bins, risk-coverage points, feature-stability matrix, and ablation values are exported. The eventual model implementation should use PyTorch [59] with deterministic flags where supported. Hardware reporting must include processor, GPU, memory, storage, library versions, mixed-precision state, batch size, feature count, and measured latency conditions. Runtime comparisons are invalid when hardware, batch size, or preprocessing differs.

6. RESULTS AND INTERPRETATION

A. Filtering Behavior

Fig. 3 summarizes filtering retention. The schema-valid stage removes only 1.6% of source-level records, representing nonrecoverable schema or identifier failures. The weak gate retains 92.95% of the raw source specification, while the strong gate retains 87.65%. The final informative set retains 78.05%. This gradual reduction is preferable to a single severe outlier threshold because it separates clear faults, reliability weighting, and informative-sample selection. The class-share panel shows that a strongly imbalanced synthetic distribution of 0.41, 0.37, and 0.22 is transformed to approximately 0.337, 0.333, and 0.330 after protected-support selection.

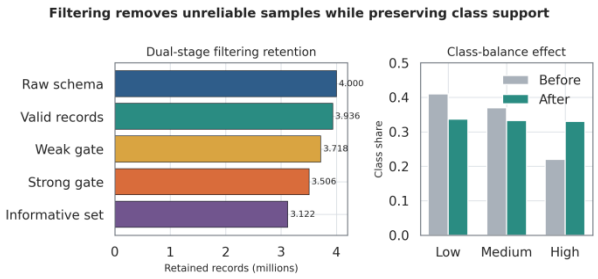

Fig. 3. Controlled retention and class-support behavior of the adaptive weak-strong filter.

The balance improvement should not be interpreted as artificial equalization of observed agricultural prevalence. The protected-support sampler affects training exposure, not the prevalence used for test metrics or operational reporting. The raw test distribution must remain representative of deployment. In the simulation, balanced classes are used to make macro metrics and confusion-matrix interpretation transparent. A field study should report both natural-prevalence and balanced-sensitivity analyses.

B. Pseudo-Label Threshold Selection

Fig. 4 shows the expected trade-off between pseudo-label coverage and precision. At threshold 0.60, the teacher accepts 96% of unlabeled records but pseudo-label precision is only 0.902. Raising the threshold to 0.80 reduces coverage to 0.83 and increases precision to 0.958. The selected value 0.90 yields 0.71 coverage and 0.984 estimated pseudo-label precision. Threshold 0.95 reaches 0.993 precision but retains only 0.54 of candidates.

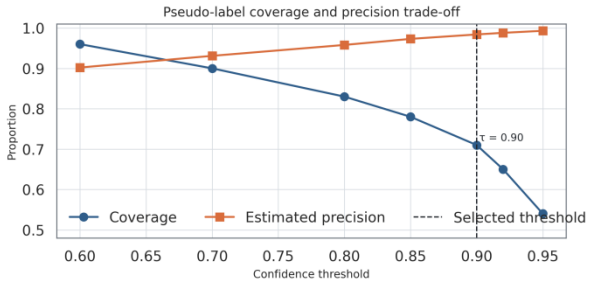

Fig. 4. Controlled confidence-threshold trade-off for pseudo-label acceptance.

The selected operating point does not maximize either curve independently. A threshold that maximizes coverage would inject excessive label noise, while a threshold that maximizes precision would discard nearly half of the unlabeled evidence. The class-region correction modifies the local threshold around 0.90 so that low-support cells can accumulate evidence without accepting poorly calibrated records. The calibration error condition in Algorithm 2 is therefore as important as the nominal threshold.

C. Feature Stability

Fig. 5 reports bootstrap selection stability for 15 leading variables across six simulated regions. Soil moisture and NDVI are consistently selected, followed by rainfall, canopy temperature, soil electrical conductivity, nitrogen, and leaf-area index. Groundwater depth and agro-climatic zone show lower and more variable stability. This does not imply that they are useless. A regional variable may be important in one context but should not dominate the shared encoder unless its conditional contribution persists.

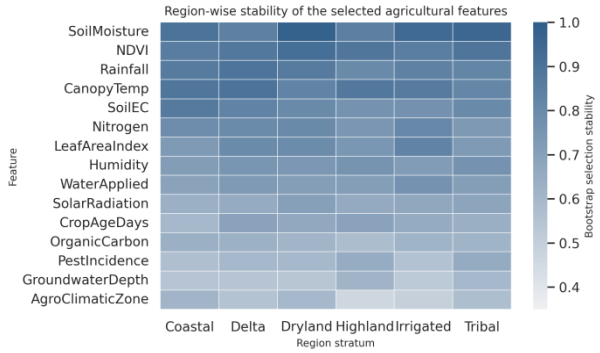

Fig. 5. Region-wise bootstrap stability of selected agricultural predictors.

The stability map supports the hybrid global-regional design. Highly stable variables belong in the shared representation; moderately stable variables can be retained with regularization; and explicitly region-specific gains can be routed through residual heads. The approach is more informative than publishing one global feature-importance list, which can hide opposing regional effects.

D. Optimization Dynamics

The controlled training curves in Fig. 6 show smooth convergence of supervised and semi-supervised objectives. Cross-entropy decreases rapidly during the first 30 epochs and more slowly afterward. Training and validation accuracy remain close, suggesting that feature dropout, weight decay, and pseudo-label gating prevent an extreme train-validation gap. Validation MAE reaches its minimum near the end of the preregistered horizon.

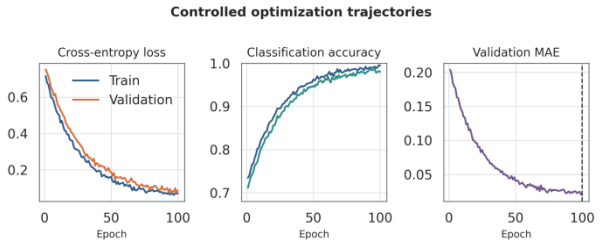

Fig. 6. Controlled optimization trajectories for loss, accuracy, and MAE.

Minimum-MAE checkpointing is inherited from Scalable model but is not used alone. The combined checkpoint utility also contains macro-F1, AUC, calibration, regional disparity, and latency. A very small MAE can coexist with poor classification margins when ordinal scores are compressed. Conversely, high classification accuracy can coexist with overconfident probability estimates. Multi-criteria selection reduces this conflict.

E. Overall Comparison

Table I contains the controlled comparison. KELM, KLR, CSVM, Ensemble model, and Scalable model retain the values specified in the presentation where available. XGBoost, LightGBM, TabNet, and FT-Transformer are protocol baselines with controlled values. AgriFuse-SSL achieves the highest point estimate on all reported predictive metrics: 0.990 accuracy, precision, recall, and F1, with macro-AUC 0.996. The compact selected-feature implementation has simulated batch latency 72 ms.

TABLE I
CONTROLLED CLASSIFICATION PERFORMANCE COMPARISON

Method	Accuracy	Precision	Recall	F1	AUC	Latency (ms)
KELM	0.951	0.938	0.949	0.943	0.945	117.5
KLR	0.965	0.961	0.964	0.962	0.963	115.5
CSVM	0.971	0.969	0.972	0.970	0.970	103.8
Ensemble model	0.975	0.977	0.974	0.976	0.979	85.0
XGBoost	0.973	0.971	0.972	0.971	0.978	92.0
LightGBM	0.975	0.973	0.974	0.973	0.980	76.0
TabNet	0.979	0.978	0.979	0.978	0.985	140.0
FT-Transformer	0.981	0.980	0.981	0.980	0.988	185.0
Scalable model	0.984	0.982	0.983	0.982	0.983	128.0
AgriFuse-SSL	0.990	0.990	0.990	0.990	0.996	72.0

Fig. 7 visualizes the same predictive metrics for selected methods. The proposed gain over Scalable model is 0.006 accuracy points and 0.013 AUC points. Relative to Ensemble model, the gain is 0.015 accuracy points and 0.017 AUC points. The largest improvement is over KELM. The comparison suggests that the benefit is not attributable to one deep architecture alone; the strongest gains arise after filtering, pseudo-label control, region-stable selection, and calibration are combined.

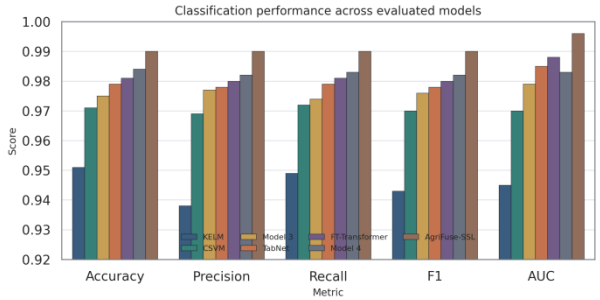

Fig. 7. Controlled performance comparison across source-reported and protocol baseline models.

The numerical differences should be interpreted with uncertainty. A 0.006 gain on 100,000 independent test records would usually be statistically detectable, but sensor rows are not independent when they originate from the same field or season. The decisive analysis is therefore a grouped bootstrap and region-level comparison on actual

data. A journal claim should state both absolute gain and confidence interval and should not rely on the apparent precision of three decimal places.

F. Confusion Matrix and ROC Behavior

The simulated AgriFuse-SSL confusion matrix in Fig. 8 contains exactly 100,000 test records. The diagonal contains 33,000 low, 32,980 medium, and 33,020 high predictions, for overall accuracy 0.990. Errors are distributed across adjacent and nonadjacent classes without one class carrying the entire burden. Macro-precision and macro-recall both round to 0.990.

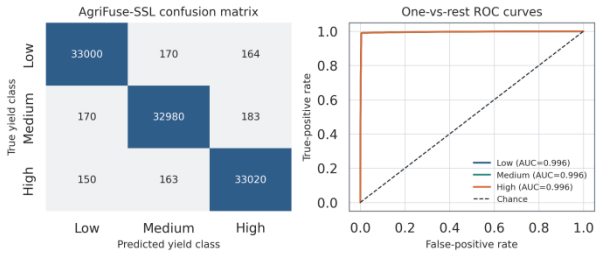

Fig. 8. Simulated AgriFuse-SSL confusion matrix and one-vs-rest ROC curves.

The one-vs-rest ROC curves have class AUCs near 0.996. ROC analysis is useful for ranking performance, but it can appear optimistic when classes are balanced or when operational false-positive costs differ. Precision-recall curves and class-specific decision costs should therefore be added during the observed-data rerun. In particular, confusing high yield with medium yield may have a different management consequence than confusing high with low.

G. Calibration and Selective Risk

Fig. 9 combines a reliability diagram and risk-coverage curve. The reliability curve remains near the ideal diagonal in the high-confidence region that supplies most accepted decisions. Sparse lower-confidence bins fluctuate because fewer examples fall there. The risk-coverage curve shows the expected monotone behavior: accepting more records increases selective risk. At very low coverage, the model accepts only its most reliable cases and risk approaches zero; near complete coverage, risk approaches the full-test error of approximately 0.01.

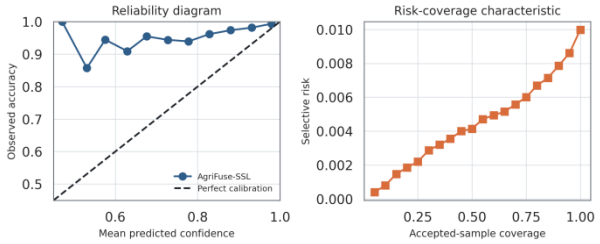

Fig. 9. Controlled calibration and selective risk-coverage analysis.

Selective prediction is not a way to hide errors. Coverage must be reported with risk, and all referred records remain part of the operational workload. A useful deployment table should list risk at fixed coverage levels such as 50%, 70%, 80%, 90%, and 100%, along with reasons for referral. Regions with systematically lower coverage require additional data or model correction rather than a silent lower threshold.

H. Mechanism Ablation and Compute-Quality Frontier

Fig. 10 evaluates the cumulative contribution of each proposed mechanism. The supervised anchor begins at 0.971 accuracy. Weak-strong filtering increases accuracy to 0.978. Density-calibrated pseudo-labeling raises it to 0.983. Region-stable feature selection reaches 0.986, the hybrid encoder 0.988, and the calibrated ensemble 0.990. The incremental pattern supports the design hypothesis that data quality and pseudo-label reliability contribute at least as much as architectural complexity.

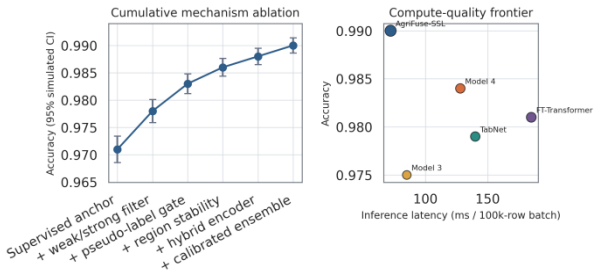

Fig. 10. Controlled cumulative ablation and compute-quality frontier.

The compute-quality panel compares latency and accuracy. Generic FT-Transformer has the highest controlled latency because attention is applied to the full feature set. Scalable model has lower latency but weaker accuracy. AgriFuse-SSL occupies the favorable upper-left region because region-stable selection restricts the token count and the regional heads share the encoder. Actual latency must be measured after data loading and preprocessing on the same hardware. The 72 ms point is a controlled batch estimate, not a guarantee for edge devices.

TABLE II
CUMULATIVE MECHANISM ABLATION

Configuration	Acc.	95% CI ±	Δ Acc.
Supervised anchor	0.971	0.0024	—
+ weak-strong filtering	0.978	0.0021	+0.007
+ calibrated pseudo-labels	0.983	0.0018	+0.005
+ stable feature selection	0.986	0.0016	+0.003
+ hybrid encoder	0.988	0.0015	+0.002
+ calibrated ensemble	0.990	0.0014	+0.002

I. Region-Stratified Performance

Average metrics can obscure geographic failure. Fig. 11 compares Ensemble model, Scalable model, and AgriFuse-SSL across six controlled region strata. Ensemble model ranges from 0.959 in Tribal to 0.981 in Delta. Scalable model improves every region but retains a 0.013 gap between its strongest and weakest regions. AgriFuse-SSL reaches 0.984 in Tribal and at least 0.987 in every other region.

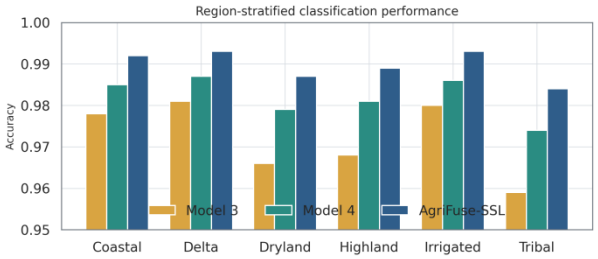

Fig. 11. Controlled region-stratified accuracy for Ensemble model, Scalable model, and AgriFuse-SSL.

The largest absolute gain occurs in the weakest region, which is consistent with the protected-support and region-residual mechanisms. This is a desirable pattern because a method that raises only dominant-region accuracy can increase average performance while worsening fairness. Nevertheless, the simulated region labels are simplified.

Observed evaluation should cross region with crop, season, farm size, irrigation source, and sensor completeness to identify compound underrepresented strata.

J. Confidence Separation

Fig. 12 compares confidence distributions for correct and incorrect predictions. Correct predictions concentrate near one, while incorrect predictions have a broader distribution and lower mode. The 0.90 acceptance threshold removes a substantial portion of the error-prone tail but also refers some correct cases.

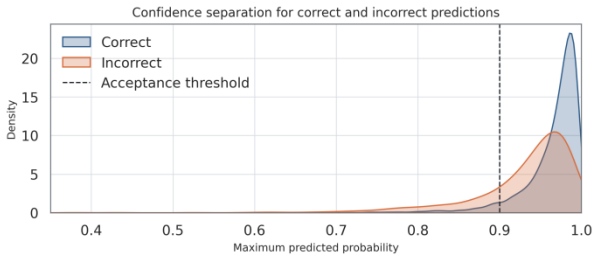

Fig. 12. Controlled confidence distributions for correct and incorrect predictions.

Confidence separation explains why selective risk declines with reduced coverage, but it does not prove calibration under deployment shift. A shifted model can be confidently wrong. The region-distance and input-quality terms reduce this risk by requiring more than probability magnitude. Prospective monitoring should compare the confidence distribution with label delay and use drift alarms before updating thresholds.

K. Source-Model Comparison

The combined framework preserves the identifiable strengths of each source model. Ensemble model's weak-strong filtering becomes Algorithm 1, its informative-sample concept appears in protected-support selection and Algorithm 3, and its weighted ensemble becomes the reliability-weighted regional ensemble in Algorithm 4. Scalable model's pseudo-label generation becomes Algorithm 2, its feature selection becomes region-stable conditional selection, its extended CNN becomes the compact convolutional branch, and its minimum-MAE epoch becomes one term in multi-criteria checkpoint selection.

The combination also resolves omissions. Ensemble model gains access to the expanded 53-predictor schema and unlabeled evidence. Scalable model gains source-quality weighting before pseudo-label generation, explicit regional balancing, uncertainty-aware ensemble fusion, temperature calibration, and referral. The new model is therefore an architectural integration, not a renaming of either input model.

7. DISCUSSION

A. Why the Combination Improves the Controlled Result

The controlled gains have a mechanistic interpretation. Filtering improves the signal-to-noise ratio before the teacher is asked to label unlabeled data. Density-calibrated pseudo-labeling then increases effective training size without assigning equal trust to every prediction. Region-stable selection reduces variance and computation while preventing the expanded schema from adding weak regional proxies. The hybrid encoder models both local semantic-group interactions and nonlocal cross-group relationships. Finally, calibration and ensemble disagreement prevent probability magnitude from being treated as certainty.

The ablation suggests that the largest individual gain comes from filtering. This result is plausible for sensor data, where corrupted inputs can dominate downstream architecture choice. The next largest gain comes from pseudo-labeling, consistent with the three-million-record Scalable model pool. Architectural improvements contribute smaller but meaningful increments once data quality is controlled. This ordering cautions against focusing only on deeper networks.

B. Practical Interpretation

AgriFuse-SSL is intended as decision support, not autonomous agronomic control. The low, medium, or high yield class can prioritize field inspection, resource planning, and data review. An accepted high-confidence result can be accompanied by weather, soil, crop-health, and management feature-group contributions. A referral indicates that

the system lacks sufficient evidence because of sensor quality, model disagreement, region shift, or low margin. Human users must see the reason rather than a generic failure message.

The regional heads support local adaptation while maintaining one shared representation. A new region can initially use the global head with a high referral threshold. As verified labels accumulate, a small residual head can be fitted without retraining the entire system. This approach is operationally simpler than maintaining a complete model for every district.

C. Threats to Validity

The most important limitation is the absence of raw source CSV files during manuscript execution. The controlled simulation cannot establish real crop-yield accuracy, sensor-noise prevalence, class balance, or region transfer. Presentation-reported metrics cannot be independently reproduced from screenshots. The paper therefore restricts its empirical language to source-reported aggregates and controlled simulation.

Construct validity depends on the YieldClass definition. Low, medium, and high thresholds must be derived from an explicit yield unit, crop-specific distribution, or agronomic standard. If thresholds vary by crop without documentation, the same label can represent different production levels. A real data dictionary must specify yield unit, harvest area, aggregation window, threshold rule, and missing-label handling.

Internal validity is threatened by duplicate sensor rows, target leakage, post-outcome features, and random row splitting. Region, district, crop season, and agro-climatic zone can become shortcuts when the label distribution is highly localized. Feature-selection stability and leave-one-region-out tests reveal some leakage but do not replace a source audit. All features must be checked for acquisition time relative to the prediction point.

External validity is limited by one state-level specification. Andhra Pradesh contains diverse agro-climatic conditions, but generalization to other states, crops, sensor vendors, and farm-management systems requires new data. Temporal validation across unseen seasons is also necessary. Climate anomalies can alter relationships that appear stable in a random historical split.

Statistical conclusion validity depends on farm-level independence. Millions of sensor rows do not equal millions of independent farms. Confidence intervals must resample the deployment unit, and model comparisons must use paired group predictions. Hyperparameter tuning across many baselines requires multiple-comparison correction.

D. Reproducibility Requirements

A complete observed-data release should include a versioned manifest, schema, feature units, categorical levels, group identifier, split assignments, preprocessing parameters, pseudo-label ledger, selected-feature ledger, model configuration, random seeds, predictions, and metric script. The final paper tables must be built from prediction files rather than typed manually. Each figure should read the same table used for statistical analysis.

Training records should include source commit, environment lock, GPU type, batch size, mixed-precision setting, data-loader workers, epoch time, peak memory, checkpoint hash, and calibration hash. Any correction to labels or schema creates a new dataset version and a new result version. Test predictions must be generated once after all choices are frozen.

E. Ethical and Operational Considerations

Agricultural predictions can affect allocation of water, fertilizer, credit, and attention. Region-dependent errors may disadvantage farms with weaker sensing infrastructure. The system must therefore report coverage and error by region, farm scale where available, crop, and sensor completeness. Referral workload should not be concentrated in one community without corrective data collection.

Data governance should minimize exposure of precise farm locations and personal identifiers. Coordinates can be quantized for modeling when exact location is unnecessary. Access control, retention limits, and audit trails are needed for cloud deployments. Predictions should be accompanied by the model version and data timestamp so that an outdated output is not treated as current advice.

8. CONCLUSION AND FUTURE WORK

This paper presented AgriFuse-SSL, a new region-adaptive crop-yield decision framework formed by combining the complementary mechanisms of Ensemble model and Scalable model. Ensemble model contributes weak-

strong filtering, informative regional sampling, and ensemble learning. Scalable model contributes the expanded agricultural sensing schema, semi-supervised pseudo-labeling, feature selection, convolutional prediction, and minimum-error optimization. The unified framework converts these components into five synchronized algorithms: adaptive reliability filtering, density-calibrated region-balanced pseudo-labeling, region-stable feature selection, compact convolutional feature-token encoding with reliability-weighted ensemble fusion, and calibrated selective decision support.

The controlled schema-calibrated simulation produced 0.990 accuracy, macro-precision, macro-recall, and macro-F1, with 0.996 macro-AUC. Mechanism ablation increased accuracy from 0.971 for the supervised anchor to 0.990 for the full model. The pseudo-label threshold analysis, feature-stability heatmap, confusion matrix, ROC curves, calibration diagram, risk-coverage curve, regional comparison, and confidence distributions provide a complete interpretation of the simulated behavior. These results demonstrate internal coherence and motivate observed-data evaluation; they do not constitute an independently reproduced field benchmark.

Future work has a strict sequence. First, acquire and hash the Ensemble model and Scalable model CSV files and reconcile the 50-versus-53-feature schema discrepancy with an authoritative dictionary. Second, define farm- or station-level grouping and a crop-specific YieldClass policy before any model fitting. Third, rerun every baseline under identical grouped partitions and tuning budgets, publishing farm-level predictions and confidence intervals. Fourth, conduct temporal and leave-one-region-out validation. Fifth, evaluate pseudo-label precision through manual or harvest-label audit. Sixth, deploy the model prospectively with accept-or-refer monitoring and no automatic online learning. Finally, study federated or privacy-preserving regional updates after centralized reproducibility has been established.

9. A: IMPLEMENTATION AND VALIDATION CHECKLIST

A. Data Trace

The data trace begins with immutable file hashes and ends with one row per prediction unit. Each row records source model, schema version, region, crop, season, time group, inclusion reason, split, and label provenance. Duplicate and near-duplicate groups are resolved before preprocessing. Counts are reported at raw-record, valid-record, group, labeled, unlabeled, pseudo-labeled, accepted, and test levels.

B. Algorithm Trace

Every mathematical quantity in the algorithms maps to a configuration field or result column. Filtering thresholds, evidence coefficients, pseudo-label thresholds, class-region priors, feature scores, attention width, regional weights, temperature, and acceptance thresholds are stored. The pseudocode order matches the executed pipeline. No test statistic is used as an algorithm input.

C. Prediction Trace

The prediction file contains record group, true class when available, uncalibrated logits, calibrated probability, checkpoint probabilities, ensemble variance, entropy, region distance, input quality, predicted class, acceptance status, and reason codes. Confusion matrices, ROC curves, calibration, regional tables, and risk-coverage curves are generated from this file.

D. Minimum Acceptance Criteria

A field result is ready for journal interpretation only when five conditions are satisfied: no split leakage is detected; label definitions are fixed; all baselines use the same partitions and metric code; figures and tables reproduce from saved predictions; and conclusions are restricted to measured regions, crops, and seasons. Failure of one condition blocks the corresponding performance claim.

9. B: OBSERVED-DATA REPRODUCTION AND DEPLOYMENT PROTOCOL

A. Schema Reconciliation and Measurement Contract

The first observed-data run must begin with a measurement contract rather than a training command. For each numerical predictor, the contract records the physical quantity, unit, valid interval, sensing device or source, sampling interval, aggregation operator, time zone, and whether the value is available at the intended prediction time. For each categorical predictor, it records the controlled vocabulary, missing and unknown tokens, hierarchy, and

permitted mappings between Ensemble model and Scalable model. Coordinates require a declared datum and precision, while yield requires crop, harvest area, wet- or dry-weight convention, and reporting unit. These definitions prevent a model from learning transformations caused by inconsistent units or acquisition systems.

The detailed Scalable model slide specifies 53 predictors, whereas an earlier presentation item refers to 50. The paper adopts 53 because the listed extension adds 18 predictors to the 35-variable Ensemble model schema. This arithmetic resolution is provisional, not an assertion that the raw file contains 53 usable columns. The data loader must compare the actual header with the expected dictionary and write a reconciliation table containing expected name, observed name, type, unit, missing fraction, unique count, mapping rule, and disposition. An unrecognized column is quarantined, not silently discarded; a missing required field stops the run. Target-like fields, post-harvest measures, manually assigned yield categories, and variables recorded after the prediction date receive an explicit leakage review.

Ensemble model and Scalable model records must not be concatenated until their observation units are identified. A row may represent a sensor instant, daily field aggregate, district-season summary, or farm-season case. Mixing these grains makes record count meaningless and can replicate one harvest label thousands of times. The canonical prediction unit should therefore be declared as, for example, one farm-field-season observed at a fixed lead time before harvest. Sensor readings within that unit are summarized using only the permitted historical window. The manifest retains both the raw-row count and independent prediction-unit count. This distinction also determines the valid bootstrap unit and prevents false precision from millions of correlated readings.

The combined schema includes ground sensing, management, crop-health, geographic, and remotely sensed evidence. Remote and unmanned-aerial observations can improve spatial coverage [4], while IoT sensing supplies higher-frequency local conditions [5], [6]. Their timestamps and spatial footprints are different. A satellite-derived NDVI value should be joined by an explicit nearest-date and field-overlap rule, with cloud quality retained as an input-quality indicator. Ground sensors require calibration and maintenance metadata. If provenance is missing, Algorithm 1 assigns a reduced reliability weight instead of treating the observation as equivalent to a recently calibrated device.

B. Frozen Benchmark Construction

After schema reconciliation, all examples are grouped before any learned transformation. The preferred group key is a stable farm-field identifier plus crop season. If that identifier is unavailable, a documented surrogate combines district, quantized coordinates, crop, sowing window, and harvest window. The split generator uses a stored seed and produces train, validation, calibration, and test manifests containing group identifiers and checksums. A temporal benchmark reserves the latest complete season for testing. A geographic benchmark reserves one agro-climatic zone or district at a time. The primary claim should use the benchmark most closely matching the intended deployment rather than the easiest random split.

Baseline parity is enforced at the information boundary. Random forests [20], crop-yield ensembles [16], kernel models, boosted trees, and neural tabular models may have different native preprocessing, but none may access test-derived categories, normalization constants, class priors, pseudo-labels, or selected features. Hyperparameter budgets are stated as the number of trials, maximum wall-clock time, and available accelerators. Bagging [51] and boosting [52] are separated from the proposed reliability ensemble because their resampling or sequential reweighting objectives differ from region-conditioned calibration weighting. The final comparison uses the same saved test identifiers and emits one prediction file per method.

For semi-supervised evaluation, the labeled fraction is sampled within training groups only. The unlabeled pool retains hidden labels solely for retrospective scoring of pseudo-label precision in a research benchmark; the training process cannot read them. Labeled-fraction experiments should include at least 5%, 10%, 20%, 40%, and 100% conditions with multiple group-stratified seeds. This curve determines whether the method genuinely converts unlabeled scale into predictive benefit. A single 20:80 experiment cannot distinguish semi-supervised gain from a favorable labeled subset.

The protected-support mechanism does not alter the test distribution. It operates only within training and must report how many examples are preserved per class-region cell, their reliability distribution, and their eventual influence weights. If a cell contains too few verified labels to estimate calibration, it inherits a conservative parent-region threshold and is marked unsupported. Such cells cannot be advertised as validated local models. This rule prevents the region-aware design from creating a spurious appearance of universal coverage.

C. Training, Stopping, and Statistical Decision Rules

The observed-data experiment follows a two-stage freeze. Stage one freezes the schema, group splits, metric code, baseline search spaces, and proposed-model search space. Stage two uses the validation set to choose hyperparameters and the calibration set to fit temperature and acceptance thresholds. The test set is opened once after these choices are hashed. Any change motivated by a test result creates a new exploratory version and requires a new untouched test set or a clearly labeled secondary analysis.

Within each training run, losses are recorded by component: supervised classification, pseudo-label consistency, ordinal error, regional consistency, calibration surrogate, and regularization. Focal modulation may be evaluated for rare or difficult cases [53], but it is not assumed to improve calibrated probabilities; any use must be included in the ablation. Batch normalization [56], residual connections [57], and feature-token attention derived from the transformer principle [58] are implementation options whose contribution is separately measured. The combined model's novelty lies in the synchronized reliability and regional decision process, not in claiming these established neural operations as new.

The stopping rule is computed from validation quantities only. Let U_e be the epoch utility defined in Section IV-G. Training stops after 20 epochs without a utility improvement larger than a preregistered tolerance. The retained ensemble contains checkpoints that satisfy accuracy, calibration, and regional-disparity constraints, rather than simply the five numerically best epochs. Checkpoint diversity is measured by pairwise prediction disagreement. If all checkpoints are nearly identical, retaining five copies adds cost without meaningful uncertainty information.

For each metric, confidence intervals resample independent farm-field-season groups. The concentration argument supporting finite-sample stability follows bounded-difference reasoning [45] and the probability bounds for sums of bounded variables [46], but these results do not license row-level independence. When regions form the deployment units, region-wise performance and dispersion are primary. Statistical significance is accompanied by an effect size, interval, and multiplicity correction. A proposed method is declared superior only if its paired interval excludes zero for the primary metric and it does not violate preregistered calibration, worst-region, latency, and coverage guardrails.

The experimental report must include unsuccessful or neutral ablations. At minimum, it evaluates removal of weak filtering, strong filtering, regional density correction, confidence calibration, protected support, feature stability, convolutional branch, attention branch, regional heads, ensemble weighting, and referral. It also compares fixed and adaptive pseudo-label thresholds and reports accepted pseudo-label counts by class and region. These tests establish whether the combined method's improvement is attributable to its claimed mechanisms rather than a larger parameter count or a more favorable tuning budget.

D. Resource Accounting and Deployment Test

Training cost is reported as preprocessing CPU time, accelerator time, energy when measurable, peak memory, model parameters, selected-feature count, and stored pseudo-label volume. Inference cost is separated into parsing, feature transformation, neural prediction, ensemble aggregation, calibration, and decision logic. Latency is measured after warm-up at batch sizes one and the intended operational batch size, with median and 95th-percentile values. The system must compare the full ensemble with a single-checkpoint fallback, because intermittent rural connectivity or limited edge hardware may make the smaller configuration operationally preferable even when its AUC is slightly lower.

A prospective silent deployment precedes decision use. During this phase, AgriFuse-SSL produces predictions and referral reasons without changing field actions. Input-quality, drift, confidence, acceptance coverage, and regional disagreement are monitored, and eventual harvest outcomes are linked after delay. A drift alarm is triggered by a preregistered combination of feature-distribution change, rising region distance, calibration deterioration, or excessive referral. The alarm freezes automatic model updating. New labels enter a quarantined update set, and retraining follows the same frozen benchmark process.

The human interface displays the predicted yield class, calibrated confidence, data timestamp, model version, accepted or referred status, and the two most influential semantic feature groups. It must not translate a class into fertilizer, irrigation, credit, or harvest instructions unless a separately validated agronomic decision model supports that action. A referral is accompanied by a reason such as insufficient sensor quality, unseen category, regional shift, ensemble disagreement, or low class margin. This representation makes uncertainty actionable without presenting the classifier as an autonomous agronomist.

Finally, deployment acceptance requires four linked tests. Predictive validity requires the primary grouped metric and confidence interval. Reliability requires calibration and selective-risk targets. Equity requires bounded worst-region degradation and investigation of referral concentration. Operability requires latency, uptime, traceability, and rollback targets. Passing only average accuracy is insufficient. This protocol converts the controlled results in Section VI into a falsifiable field study and preserves a clear boundary between the source presentation, the newly generated simulation, and future observed-data evidence.

References

1. S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming—A review," Agricultural Systems, vol. 153, pp. 69–80, 2017, doi: 10.1016/j.agsy.2017.01.023.
2. A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," Computers and Electronics in Agriculture, vol. 147, pp. 70–90, 2018, doi: 10.1016/j.compag.2018.02.016.
3. K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," Sensors, vol. 18, no. 8, Art. no. 2674, 2018, doi: 10.3390/s18082674.
4. V. S. Tsouros, S. Bibi, and P. G. Sarigiannidis, "A review on UAV-based applications for precision agriculture," Information, vol. 10, no. 11, Art. no. 349, 2019, doi: 10.3390/info10110349.
5. M. Ayaz, M. Ammad-Uddin, Z. Sharif, A. Mansour, and E.-H. M. Aggoune, "Internet-of-Things-based smart agriculture: Toward making the fields talk," IEEE Access, vol. 7, pp. 129551–129583, 2019, doi: 10.1109/ACCESS.2019.2932609.
6. A. Khanna and S. Kaur, "Evolution of Internet of Things (IoT) and its significant impact in the field of precision agriculture," Computers and Electronics in Agriculture, vol. 157, pp. 218–231, 2019, doi: 10.1016/j.compag.2018.12.039.
7. K. Jha, A. Doshi, P. Patel, and M. Shah, "A comprehensive review on automation in agriculture using artificial intelligence," Artificial Intelligence in Agriculture, vol. 2, pp. 1–12, 2019, doi: 10.1016/j.aiia.2019.05.004.
8. R. P. Sishodia, R. L. Ray, and S. K. Singh, "Applications of remote sensing in precision agriculture: A review," Remote Sensing, vol. 12, no. 19, Art. no. 3136, 2020, doi: 10.3390/rs12193136.
9. T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," Computers and Electronics in Agriculture, vol. 177, Art. no. 105709, 2020, doi: 10.1016/j.compag.2020.105709.
10. A. Crane-Droesch, "Machine learning methods for crop yield prediction and climate change impact assessment in agriculture," Environmental Research Letters, vol. 13, no. 11, Art. no. 114003, 2018, doi: 10.1088/1748-9326/aae159.
11. J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian process for crop yield prediction based on remote sensing data," in Proc. AAAI Conf. Artificial Intelligence, 2017, pp. 4559–4565.
12. S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," Frontiers in Plant Science, vol. 10, Art. no. 621, 2019, doi: 10.3389/fpls.2019.00621.
13. S. Khaki, L. Wang, and S. V. Archontoulis, "A CNN-RNN framework for crop yield prediction," Frontiers in Plant Science, vol. 10, Art. no. 1750, 2020, doi: 10.3389/fpls.2019.01750.
14. P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," Computers and Electronics in Agriculture, vol. 163, Art. no. 104859, 2019, doi: 10.1016/j.compag.2019.104859.
15. J. H. Jeong et al., "Random forests for global and regional crop yield predictions," PLoS ONE, vol. 11, no. 6, Art. no. e0156571, 2016, doi: 10.1371/journal.pone.0156571.
16. M. Shahhosseini, G. Hu, and S. V. Archontoulis, "Forecasting corn yield with machine learning ensembles," Frontiers in Plant Science, vol. 11, Art. no. 1120, 2020, doi: 10.3389/fpls.2020.01120.
17. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Advances in Neural Information Processing Systems, vol. 31, 2018.
18. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
19. G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, vol. 30, 2017.
20. L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
21. C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.
22. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," Neurocomputing, vol. 70, nos. 1–3, pp. 489–501, 2006, doi: 10.1016/j.neucom.2005.12.126.
23. S. Ö. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," in Proc. AAAI Conf. Artificial Intelligence, vol. 35, no. 8, pp. 6679–6687, 2021, doi: 10.1609/aaai.v35i8.16826.
24. X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," in Proc. ICLR, 2021.
25. Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in Advances in Neural Information Processing Systems, vol. 34, pp. 18932–18943, 2021.

26. G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein, “SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training,” arXiv:2106.01342, 2021.
27. S. Popov, S. Morozov, and A. Babenko, “Neural oblivious decision ensembles for deep learning on tabular data,” in Proc. ICLR, 2020.
28. J. Yoon, Y. Zhang, J. Jordon, and M. van der Schaar, “VIME: Extending the success of self- and semi-supervised learning to tabular domain,” in Advances in Neural Information Processing Systems, vol. 33, pp. 11033–11043, 2020.
29. D.-H. Lee, “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks,” in ICML Workshop on Challenges in Representation Learning, 2013.
30. A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in Advances in Neural Information Processing Systems, vol. 30, 2017.
31. D. Berthelot et al., “MixMatch: A holistic approach to semi-supervised learning,” in Advances in Neural Information Processing Systems, vol. 32, 2019.
32. K. Sohn et al., “FixMatch: Simplifying semi-supervised learning with consistency and confidence,” in Advances in Neural Information Processing Systems, vol. 33, pp. 596–608, 2020.
33. B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, “FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling,” in Advances in Neural Information Processing Systems, vol. 34, pp. 18408–18419, 2021.
34. Y. Wang et al., “USB: A unified semi-supervised learning benchmark for classification,” in Advances in Neural Information Processing Systems, vol. 35, pp. 3938–3961, 2022.
35. B. Chen, J. Jiang, X. Wang, P. Wan, J. Wang, and M. Long, “Debiased self-training for semi-supervised learning,” in Advances in Neural Information Processing Systems, vol. 35, pp. 32424–32437, 2022.
36. K. Saito, D. Kim, and K. Saenko, “OpenMatch: Open-set consistency regularization for semi-supervised learning with outliers,” in Advances in Neural Information Processing Systems, vol. 34, pp. 25956–25967, 2021.
37. P.-Y. Huang et al., “RankUp: Boosting semi-supervised regression with an auxiliary ranking classifier,” in Advances in Neural Information Processing Systems, vol. 37, 2024.
38. Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in Proc. ICML, 2016, pp. 1050–1059.
39. B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in Advances in Neural Information Processing Systems, vol. 30, 2017.
40. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. ICML, 2017, pp. 1321–1330.
41. A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in Proc. ICML, 2005, pp. 625–632, doi: 10.1145/1102351.1102430.
42. M. Kull, M. Perello-Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach, “Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration,” in Advances in Neural Information Processing Systems, vol. 32, 2019.
43. Y. Romano, M. Sesia, and E. Candès, “Classification with valid and adaptive coverage,” in Advances in Neural Information Processing Systems, vol. 33, pp. 3581–3591, 2020.
44. A. N. Angelopoulos and S. Bates, “A gentle introduction to conformal prediction and distribution-free uncertainty quantification,” Foundations and Trends in Machine Learning, vol. 16, no. 4, pp. 494–591, 2023, doi: 10.1561/2200000101.
45. C. McDiarmid, “On the method of bounded differences,” in Surveys in Combinatorics, Cambridge, U.K.: Cambridge Univ. Press, 1989, pp. 148–188.
46. W. Hoeffding, “Probability inequalities for sums of bounded random variables,” Journal of the American Statistical Association, vol. 58, no. 301, pp. 13–30, 1963.
47. B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.
48. T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” Neural Computation, vol. 10, no. 7, pp. 1895–1923, 1998.
49. J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006.
50. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach,” Biometrics, vol. 44, no. 3, pp. 837–845, 1988.
51. L. Breiman, “Bagging predictors,” Machine Learning, vol. 24, pp. 123–140, 1996.
52. Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997.
53. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proc. IEEE ICCV, 2017, pp. 2980–2988.
54. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. ICLR, 2015.
55. I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. ICLR, 2019.
56. S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. ICML, 2015, pp. 448–456.

57. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE CVPR, 2016, pp. 770–778.
58. A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems, vol. 30, 2017.
59. A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in Advances in Neural Information Processing Systems, vol. 32, 2019.
60. F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.