ISNNI: 3007-3391 HEC Category: Y

Journal of Computers and Intelligent Systems
Vol. 4, No. 2, 2026, pages 67 - 81
Journal Homepage:
https://journals.iub.edu.pk/index.php/JCIS/

Cross-Scale Mixture-of-Experts with Conformal Uncertainty for Reliable Crop Yield Prediction across Latent Agro-Climate Regimes

Ayesha Qureshi[1], Abdul Haseeb Wajid[1], Ali Samad[1],
[1]Affilition of Author 1, The Islamia University Bahawalpur.

ARTICLE INFO

Article History:

Received:	January	18, 2026
Revised:	April	16, 2026
Accepted:	May	20, 2026
Available Online:	June	30, 2026

Keywords:

Crop yield prediction
Mixture-of-experts (MoE)
Agro-climatic regimes
Unsupervised clustering
Conformal prediction
Uncertainty quantification
Cross-scale modeling
Tabular deep learning
Precision agriculture

Classification Codes:

Funding:

This research received no specific grant from any funding agency in the public or not-for-profit sector.

ABSTRACT

Accurate and reliable crop-yield prediction is foundational for climate-resilient agricultural planning, yet data-driven models are commonly developed either for micro plot-scale records or macro country-level panels in isolation and often without decision-ready uncertainty. This paper presents AgroX-MoE-CP, a cross-scale mixture-of-experts (MoE) framework equipped with split conformal prediction to deliver calibrated prediction intervals across heterogeneous agro-climatic regimes. First, latent agro-climate regimes (AgroClusters) are discovered via unsupervised clustering of standardized weather and management variables, capturing heterogeneity without manual zoning. AgroClusters are then provided as an additional categorical context to a tabular MoE model, where a gating network adaptively blends expert predictors. On a synthetic plot-level dataset (1M records), we inject agronomy-aware engineered features (rainfall intensity and an approximate growing degree days indicator). On a FAO-based macro panel, we stabilize learning by regressing on a log-transformed yield and map predictions back to the original units. Empirically, AgroX-MoE achieves strong point accuracy at both scales (micro: $R^2 \approx 0.91$, MAE≈0.40 t/ha; macro: $R^2 \approx 0.88$, MAE≈1.68 t/ha equivalent). Split conformal calibration yields finite-sample-valid 90% prediction intervals with empirical coverage ≈0.90 on both datasets and moderate average widths (micro: ≈1.66 t/ha; macro: ≈8.88 t/ha equivalent). Results indicate that combining latent regime discovery, MoE specialization, agronomic feature engineering, and distribution-free conformal calibration supports accurate and reliable crop-yield forecasting across scales.

Corresponding Author's Email:

Citation:

1. Introduction

Climate variability and extremes increasingly disrupt agricultural production, placing pressure on food systems and planning under uncertainty. Yield forecasting supports interventions across the management lifecycle input allocation, irrigation scheduling, insurance pricing, and supply-chain risk mitigation. Recent reviews highlight rapid growth in machine-learning and deep-learning approaches for yield prediction, driven by increasing availability of agronomic, weather, and management covariates as well as large-scale remote sensing products [2].

Despite progress, two persistent gaps limit decision readiness. First, many models target a single scale: (i) micro plot/field-scale datasets used for precision agriculture and management optimization, or (ii) macro country year panels used for policy evaluation and long-term food-security planning. Models optimized at one scale may not transfer to the other because the relevant drivers, noise characteristics, and aggregation effects differ. Second, uncertainty quantification is often ad hoc (e.g., assumed Gaussian residuals or uncalibrated neural-network heuristics), whereas downstream decisions require valid prediction intervals with measurable coverage [1].

A key source of predictive difficulty is heterogeneity of growing conditions. Within the same nominal administrative zone or crop type, yield responses vary due to interactions between rainfall, temperature, soil, and management. Traditional approaches often handle heterogeneity through manual agro-ecological zoning or expert-crafted strata, which may not align with the data distribution and can be brittle under climate change. We instead learn latent regimes directly from the data using unsupervised clustering of standardized covariates.

To exploit such regimes, we propose AgroX-MoE-CP: a cross-scale tabular mixture-of-experts model augmented with conformal prediction. The mixture-of-experts (MoE) paradigm provides conditional specialization: a gating network assigns each example a distribution over expert predictors, and the final prediction is a weighted combination. MoE has become a standard strategy for handling heterogeneous data distributions and scaling capacity efficiently [3]. However, specialization alone does not guarantee reliable uncertainty.

To obtain distribution-free reliability, we employ split conformal prediction, which constructs prediction intervals by calibrating residual quantiles on a held-out calibration set. Conformal prediction provides finite-sample coverage guarantees under minimal assumptions (primarily exchangeability), and has recently been studied in agricultural decision support and related domains [1], [5]. We additionally evaluate group-conditional calibration by crop type (a Mondrian-style strategy) to diagnose fairness and robustness across subpopulations.

Experiments follow a consistent cross-scale protocol on two publicly available datasets: (i) a synthetic micro plot-level dataset with 1M records and categorical management descriptors, and (ii) a FAO-derived macro panel of ≈28k country crop year records (yield in hg/ha). On micro data, we add agronomy-aware engineered features capturing rainfall intensity and a simplified growing degree days approximation. On macro data, we regress on a log-transformed yield to stabilize optimization and map predictions back to the original scale. Across both datasets, we report point accuracy (MAE, RMSE, R^2) and uncertainty quality (conformal coverage and interval width).

The contributions of this work are fourfold:

• We introduce AgroX-MoE-CP, a cross-scale tabular mixture-of-experts framework that integrates unsupervised latent regime discovery (AgroClusters) with distribution-free conformal uncertainty calibration.

• We design a practical preprocessing pipeline that combines agronomy-informed feature engineering at the micro scale with log-stabilized regression at the macro scale.

• We conduct an empirical study on two datasets spanning plot-level and country–crop–year scales, including ablations and baseline comparisons on the micro dataset, as well as uncertainty diagnostics (global and crop-wise coverage) on both datasets.

• We provide a reproducible experimental recipe including hyperparameters, data splits, and evaluation protocol to facilitate future benchmarks on cross-scale, reliability-aware crop-yield prediction.

2. Related Work

A. Crop yield prediction at micro and macro scales

Crop-yield modeling spans a spectrum from plot-level management optimization to national-scale monitoring. At micro scale, models learn yield responses to localized weather, soil, and management variables (e.g., fertilizer and irrigation decisions). At macro scale, panel datasets aggregate yield across administrative units and years and are frequently used for policy analysis and long-term projections. Comprehensive surveys summarize a diverse toolbox including linear regression, tree ensembles, deep neural networks, and hybrid crop–ML pipelines [2].

Recent macro-scale efforts increasingly adopt machine-learning models to capture nonlinear interactions between rainfall, temperature, and management proxies, often coupling FAO-derived yield series with climate reanalysis and socioeconomic drivers. However, the macro setting is challenged by measurement noise, changing reporting standards, and nonstationary driven by technology and climate trends. Log-transformations, detrending, and representation learning are commonly used to stabilize modeling under heavy-tailed yield distributions.

Across both scales, a recurring limitation is insufficient uncertainty reporting. While probabilistic neural networks, Bayesian methods, and quantile regression have been explored, their calibration can be brittle under distribution shift. Recent work in crop production decision support explicitly examines conformal prediction as a practical, distribution-free alternative that can wrap around arbitrary point models [1].

B. Deep learning for tabular agricultural data

Many agronomic datasets are fundamentally tabular: a mix of categorical descriptors (region, crop type, soil class) and numeric drivers (weather, management intensities). Deep models for tabular data typically embed categorical variables and fuse them with numeric features via multilayer perceptron's or attention-based architectures. A recent survey of deep tabular learning highlights design patterns such as feature embeddings, attention blocks, and hybrid architectures that combine transformer components with MLP backbones [4].

In practice, strong tree ensembles (e.g., gradient-boosted decision trees) remain competitive on tabular problems, motivating careful baseline comparisons and ablations. Our experiments therefore include linear regression, random forests, and XGBoost baselines on the micro dataset alongside deep models.

C. Mixture-of-experts for heterogeneous regimes

Mixture-of-experts models combine multiple specialist predictors through a gating function that assigns input-dependent weights. The approach supports conditional computation and can allocate capacity to distinct sub-regimes of the data distribution. Modern MoE variants have been extensively studied in large-scale representation learning and are surveyed in [3].

Within agronomy, MoE is appealing because yield drivers differ across agro-climatic and management regimes. For example, rainfall variability may dominate in rainfed systems, whereas temperature and irrigation constraints may dominate in irrigated systems. A gating network can learn to route examples to experts that focus on different response surfaces, potentially improving accuracy and interpretability under heterogeneity.

D. Distribution-free uncertainty with conformal prediction

Conformal prediction constructs prediction intervals with guaranteed marginal coverage under exchangeability. Split conformal prediction is especially practical: it trains a point model on a training set and calibrates residual quantiles on a held-out calibration set. A tutorial introduction and broader discussion of distribution-free uncertainty can be found in [5].

Variants include conformalized quantile regression, which combines quantile models with conformal calibration, and group-conditional (Mondrian) conformal methods that target balanced coverage across predefined subgroups. In this work, we employ a symmetric absolute-residual score and evaluate both global and crop-conditional calibration.

3. Problem Formulation and Datasets

We consider a supervised regression setting for crop yield prediction. Each example consists of mixed-type features $x \in \mathbb{R}^{d_{\text{num}}} \times \mathcal{C}^{d_{\text{cat}}}$ and a continuous target $y \in \mathbb{R}$. The objective is to learn a predictor $f(x)$ that minimizes the expected loss while also providing reliable prediction intervals for decision-making. In this paper, we evaluate the same modeling pipeline at two scales: micro-scale plot-level records and macro-scale country–crop–year panels.

Let $D_{\text{train}}, D_{\text{val}}, D_{\text{calib}}$, and D_{test} denote the training, validation, calibration, and test splits, respectively. Point models are trained on D_{train}, tuned using D_{val}, and then calibrated using D_{calib} to produce conformal prediction intervals evaluated on D_{test}. All reported test metrics are computed on D_{test} only.

A. Datasets

Table I summarizes the two datasets used in this study, including split sizes and target definitions.

Table I. Dataset summary (after preprocessing).

Dataset	Scale	Records	Train/Val/Calib/Test	Target	Inputs
Micro (Synthetic, Kaggle)	Plot-level	1000000	600k / 100k / 100k / 200k	Yield (t/ha)	4 cat + 2 bin + 3 num + 2 eng. + AgroCluster
Macro (FAO-based panel)	Country–crop–year	28242	16944 / 2824 / 2825 / 564	Yield (hg/ha) → log1p	Area, Item, Year_Bin, AgroCluster + 4 numeric

The micro dataset is a synthetic, plot-level dataset comprising 1,000,000 records, with the target yield measured in tons per hectare (t/ha). Features include categorical descriptors (Region, Soil_Type, Crop, Weather_Condition), two binary management indicators (Fertilizer_Used, Irrigation_Used), and numeric weather/growth attributes (Rainfall_mm, Temperature_Celsius, Days_to_Harvest). The large sample size enables stable training of deep models and robust calibration of prediction intervals.

The macro dataset is an FAO-derived panel (≈28k records) with the target yield reported in hectograms per hectare (hg/ha). Covariates include Year, average annual rainfall, pesticide usage, and average temperature, along with categorical identifiers for country/area and crop item. To reduce skewness and stabilize optimization, we train on log_yield = log(1 + hg/ha_yield) and map predictions back via expm1. For interpretability, we also report an equivalent conversion to tons per hectare using 1 hg/ha = 0.0001 t/ha.

4. AGROX-MOE-CP Methodology

Fig. 1 and Fig. 2 illustrate the proposed cross-scale pipeline and the AgroX-MoE architecture, respectively.

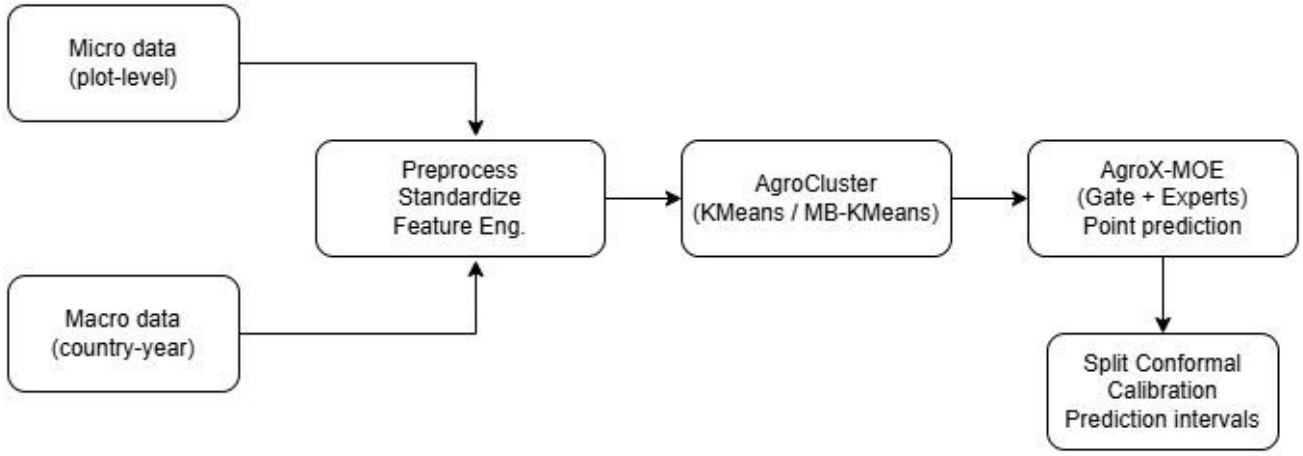

Figure. 1. *Overview of the AgroX-MoE-CP pipeline across micro and macro scales.*

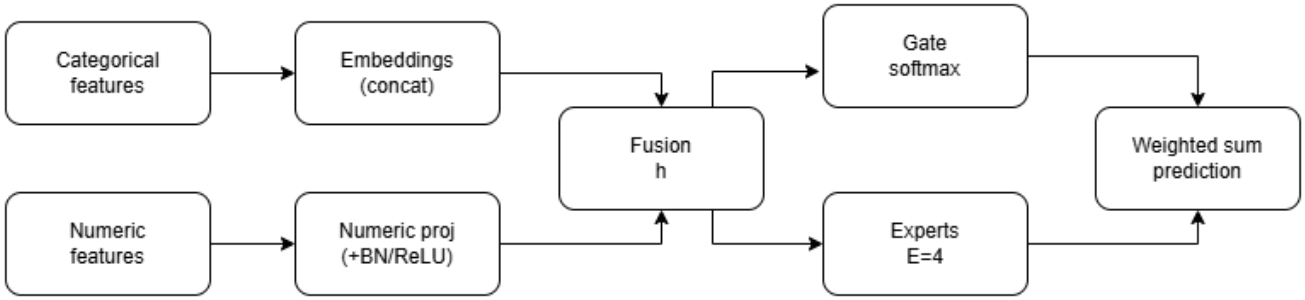

Figure. ***2.*** *AgroX-MoE architecture: categorical embeddings and numeric projection are fused; a gating network blends multiple expert predictors.*

A. Agronomy-aware feature engineering

To inject domain knowledge while keeping the model lightweight, we derive two additional micro-scale features that approximate water and thermal accumulation over the growing period.

1. Rainfall intensity. Given total rainfall R(mm) and days-to-harvest T(days), we compute a simple intensity proxy:

$$(1)\quad \text{Rain\_Intensity} = \frac{R}{\max(1, T)}.$$

2. Growing degree days (GDD) approximation. Using a common base temperature of 10 °C, we define:

$$(2)\quad \text{GDD\_Approx} = \max(0, \text{Temp} - 10) \times T.$$

These engineered features aim to capture nonlinear stress effects: the same seasonal rainfall can have different impacts depending on the length of the growing window, and yield sensitivity to temperature depends on cumulative thermal exposure.

For the macro panel, we introduce a categorical variable $\text{Year_Bin} = 5 \times \lfloor \text{Year}/5 \rfloor$to capture slow technological and climate trends at a coarse temporal resolution. We additionally model the transformed target $\text{log_yield} = \log(1 + y)$to reduce skewness and improve numerical stability.

B. Preprocessing: standardization and categorical encoding

Numeric features are standardized to have zero mean and unit variance to improve optimization and to ensure that clustering is not dominated by scale differences. On the micro dataset, we compute standardization statistics using the training split and apply the same transformation to the validation, calibration, and test splits. On the macro panel, we apply standardization as an unsupervised preprocessing step prior to clustering and modeling; while this does not use target values, it can expose test-set distributional information and is therefore discussed as a limitation.

Categorical features are mapped to integer identifiers and embedded through learned lookup tables. Binary management indicators are cast to $\{0, 1\}$and treated as numeric inputs. The final input vector for each example is a concatenation of (i) categorical embeddings and (ii) standardized numeric and binary features.

C. Latent agro-climatic regime discovery (AgroClusters)

We hypothesize that heterogeneous yield responses arise from latent agro-climatic regimes characterized by weather and management covariates. Rather than relying on manual zoning, we discover such regimes via unsupervised clustering on standardized numeric drivers.

Let $z_i \in \mathbb{R}^d$denote the standardized numeric feature vector for sample i. We fit KMeans (or MiniBatchKMeans for scalability) to obtain Kcluster centroids $\{\mu_k\}_{k=1..K}$and assign each sample an AgroCluster label $c_i \in \{1, \ldots, K\}$. The cluster assignment is then treated as an additional categorical feature.

In the micro experiments, we use $K = 4$clusters for the baseline AgroMoE study and $K = 6$clusters for the cross-scale AgroX variant. For the macro panel, we use $K = 6$to capture a moderate number of regimes without over-fragmenting the smaller dataset. We note that the clustering objective is purely unsupervised and does not directly optimize for yield prediction; the MoE model can learn whether and how to exploit the regime information.

D. Tabular mixture-of-experts predictor

The AgroX-MoE predictor is a tabular neural network with (i) categorical embeddings, (ii) a gating network, and (iii) multiple expert MLPs. Each categorical feature $j$with cardinality $C_j$is embedded into $\mathbb{R}^{e_j}$, and the embedding vectors are concatenated. Numeric features are optionally projected through a linear layer followed by a nonlinearity.

Let $e(x_{\text{cat}})$denote the concatenated embedding vector and x_{num}denote the numeric (and binary) feature vector. The fused representation is

$$(3)\quad h = [\, e(x_{\text{cat}}) \,;\, x_{\text{num}} \,].$$

The gating network outputs logits $s \in \mathbb{R}^E$that are normalized via a softmax function to produce mixture weights g:

$$(4)\ \ g = \text{softmax}(\text{Gate}(h)), \text{where } g_k \geq 0 \text{ and } \sum_k g_k = 1.$$

Each expert network $f_k(h)$outputs a scalar prediction. The final point prediction is the weighted sum

$$(5)\ \ \hat{y} = \sum_{k=1..E} g_k \cdot f_k(h).$$

Experts share the same input representation but have independent parameters, enabling specialization. In our experiments, $E = 4$experts and two-hidden-layer MLP experts are used. Smooth L1 loss (Huber loss) is employed to reduce sensitivity to heavy-tailed errors.

E. Training objective and evaluation metrics

Models are trained with mini-batch stochastic optimization. For the micro experiments, we use Adam with a learning rate of 1×10^{-3}, batch size 2048, dropout 0.2, and 10 epochs. For the macro panel, we use AdamW with weight decay 1×10^{-4}, learning rate 1×10^{-3}, batch size 512, dropout 0.1, and 10 epochs. We select the best checkpoint using validation performance (RMSE for micro, MAE for macro), then evaluate the selected model on the test data.

Point-accuracy metrics include mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2):

$$(6)\ \ \text{MAE} = \frac{1}{n}\sum_i |\, y_i - \hat{y}_i \,|,$$

$$(7)\ \ \text{RMSE} = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2},$$

$$(8)\ \ R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}.$$

All metrics are reported on the held-out test set. On the macro dataset, metrics are computed on the original yield scale after mapping log-space predictions back via expm1.

F. Split conformal prediction and crop-conditional calibration

To quantify predictive uncertainty with finite-sample guarantees, we use split conformal prediction. After training a point model $\hat{y} = f(x)$, we compute calibration residual scores on a disjoint calibration set D_{calib}:

$$(9)\ \ r_i = |\, y_i - \hat{y}_i \,|, (x_i, y_i) \in D_{\text{calib}}.$$

Given a target miscoverage level α($\alpha = 0.1$ for 90% intervals), we compute the $(1-\alpha)$-quantile of residuals:

$$(10)\ \ \hat{q} = \text{Quantile}_{1-\alpha}(\{r_i\}).$$

For any new example x, the symmetric conformal prediction interval is

$$(11)\ \ \text{PI}(x) = [\, \hat{y}(x) - \hat{q},\ \hat{y}(x) + \hat{q} \,].$$

Under exchangeability, split conformal prediction guarantees marginal coverage:

$$P(y \in \text{PI}(x)) \geq 1 - \alpha.$$

Interval efficiency is summarized by the average width $2\hat{q}$and empirical coverage on the test set.

To diagnose group-level reliability, we also compute crop-conditional quantiles $\hat{q}_c$for each crop group $c$using only calibration residuals belonging to crop c. For test points with crop label c, we use

$$\text{PI}_c(x) = [\, \hat{y}(x) - \hat{q}_c,\ \hat{y}(x) + \hat{q}_c \,],$$

with fallback to the global $\hat{q}$if a group is too small. While this Mondrian strategy does not provide full conditional coverage, it helps balance error across predefined strata and reveals whether some crops experience systematic under- or over-coverage [1], [5].

We note that symmetric absolute-residual intervals are simple and robust but may be conservative when residual distributions are skewed. Future extensions could employ conformalized quantile regression to produce asymmetric intervals while maintaining coverage [6].

5. Experimental Setup

We implement all models in PyTorch and perform clustering and baselines using scikit-learn. Experiments are executed on a GPU-enabled environment (CUDA). For reproducibility, random seeds are fixed for NumPy, Python, and PyTorch.

We follow a consistent four-way split protocol: 60% train, 10% validation, 10% calibration, and 20% test. On the micro dataset, splits are stratified by crop type to preserve crop distribution across partitions.

Table II summarizes key hyperparameters for the micro and macro experiments.

Table II. Training and model hyperparameters.

Setting	Micro	Macro
Experts (E)	4	4
Hidden dimension	64	128
Dropout	0.2	0.1
Optimizer	Adam	AdamW (wd=1e-4)
Learning rate	1e-3	1e-3
Batch size	2048	512
Epochs	10	10
Loss	Smooth L1	Smooth L1
AgroCluster K	4 (baseline) / 6 (AgroX)	6
Conformal α	0.1	0.1

A. Baselines and ablations (micro dataset)

To contextualize performance on the micro dataset, we compare AgroMoE-CP against several baselines trained on the same splits and feature set:

• Linear regression on the concatenated (label-encoded) categorical, binary, and standardized numeric features.

• Random forest regressor (200 trees, min leaf size 2).

• XGBoost regressor with 500 estimators and standard subsampling settings.

In addition, we evaluate two neural ablations: (i) AgroMLP, which removes the MoE structure and uses a single MLP backbone, and (ii) AgroMoE without AgroCluster, which evaluates whether the unsupervised regime feature adds value beyond the original covariates.

B. Uncertainty evaluation

Uncertainty is evaluated using split conformal prediction with α=0.1. We report (i) empirical coverage on the test set and (ii) average interval width. For micro data, we additionally compute crop-wise coverage using crop-conditional quantiles to assess subgroup reliability.

6. Results and Analysis

A. Micro dataset: point prediction performance

Table III reports test-set performance on the 1M-record micro dataset. All models achieve high R^2 (≈0.91) and low error (≈0.40 t/ha MAE), indicating that the synthetic data-generating process is well captured by relatively simple function

classes. AgroMoE-CP is competitive with strong baselines but does not materially outperform linear regression or gradient boosting on this dataset. This result suggests that the dominant signal may be largely additive/linear after preprocessing, and that additional capacity yields diminishing returns.

Importantly, AgroMoE-CP provides calibrated uncertainty through conformal prediction (Section VI-C), which the point baselines do not supply by default. In operational settings, calibrated intervals can be more valuable than marginal improvements in RMSE.

Table III. Micro dataset test performance (lower MAE/RMSE is better; higher R^2 is better).

Model	MAE	RMSE	R2
Linear Regression	0.400	0.501	0.912
XGBoost	0.401	0.502	0.912
AgroMoE (no AgroCluster)	0.401	0.503	0.912
AgroMoE-CP (proposed)	0.401	0.503	0.912
AgroMLP (no MoE)	0.403	0.505	0.911
Random Forest	0.411	0.515	0.907

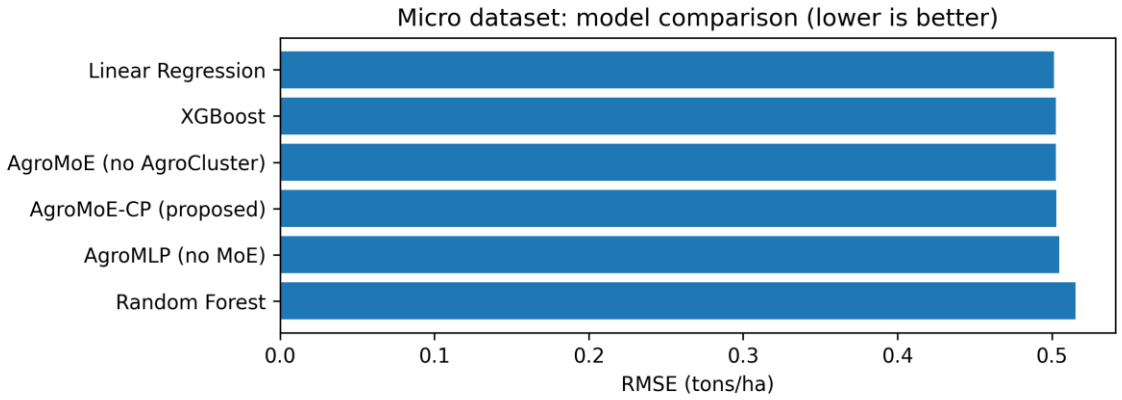

Figure. *3.* *Micro dataset: RMSE comparison across models.*

B. Learning dynamics and ablation insights (micro)

Fig. 4 shows validation RMSE across epochs for the neural models. All variants converge rapidly within a few epochs, and later epochs yield marginal changes. The no-MoE ablation (AgroMLP) performs slightly worse than the MoE variants, suggesting modest benefits from expert specialization. However, the AgroCluster feature contributes minimally: removing AgroCluster yields nearly identical RMSE, implying that the learned clusters either align with existing categorical descriptors or are not strongly predictive in the synthetic setting.

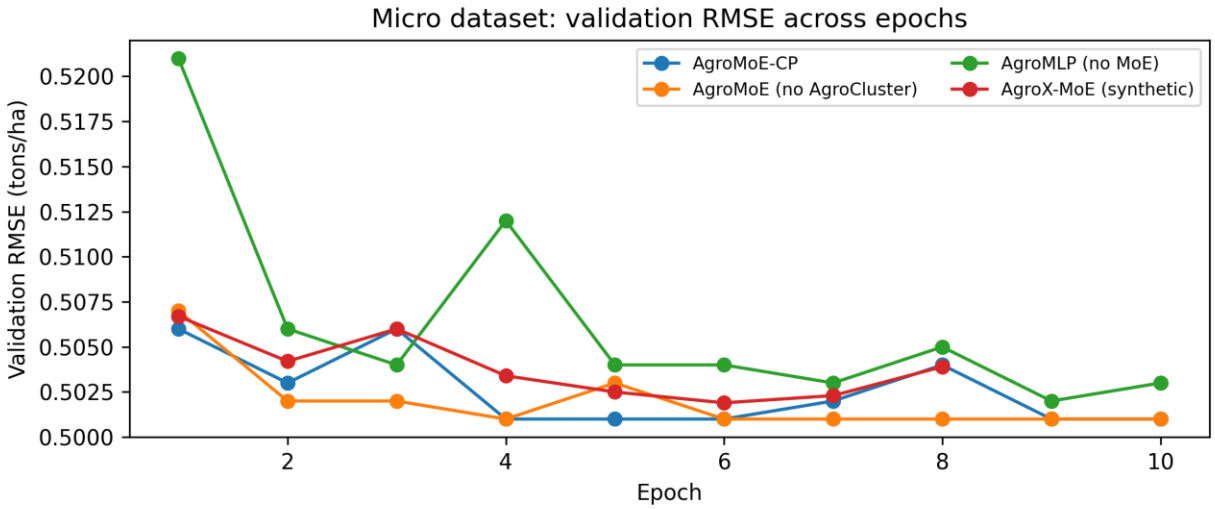

Figure. *4.* *Micro dataset: validation RMSE across epochs for neural variants.*

C. Macro dataset: point prediction performance

Table IV summarizes AgroX-MoE test performance on the FAO-derived macro panel. Training is performed in log space (log_yield $= \log(1 + y)$), and predictions are mapped back to the original yield scale for evaluation. The model achieves $R^2 \approx 0.88$and RMSE $\approx 30{,}000$hg/ha (≈3.01 t/ha equivalent). This indicates that the same regime-aware MoE design can effectively model a substantially different data regime characterized by aggregation noise and temporal variation.

Fig. 5 shows that the macro validation error decreases sharply after the first epoch and then improves steadily. The large first-epoch error is consistent with the high variance of the untrained model in log space; subsequent epochs rapidly learn the dominant country/crop effects captured by the categorical embeddings.

Table IV. Macro dataset test performance (original scale).

Dataset	MAE (hg/ha)	RMSE (hg/ha)	R2	MAE (t/ha equiv.	RMSE (t/ha equiv
Macro (FAO pane	16754.3145	30103.1582	0.8751	1.675	3.01

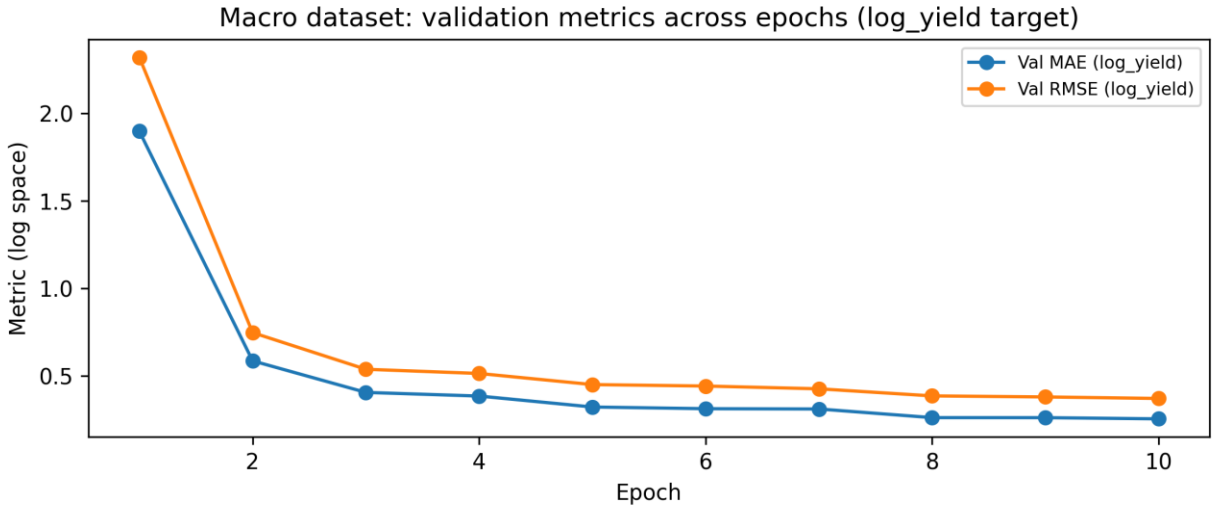

Figure. *5.* *Macro dataset: validation MAE/RMSE in log_yield space across epochs.*

D. Conformal uncertainty: coverage and efficiency

Table V reports conformal prediction interval diagnostics at both scales. For both micro and macro datasets, empirical coverage is close to the nominal 90% target (≈0.90), demonstrating effective calibration. Interval width differs substantially across scales due to differences in noise and target units: the macro panel exhibits wider intervals (≈8.88 t/ha equivalent) than the micro dataset (≈1.66 t/ha).

Fig. 6 reports crop-wise coverage for group-conditional conformal prediction on the micro dataset. Coverage is relatively uniform across crops (≈0.896–0.905), suggesting that the residual distribution is similar across crop types and that global calibration is already adequate. In settings where some crops have systematically higher uncertainty, crop-conditional calibration may improve fairness and interpretability.

Table V. Conformal prediction interval diagnostics (90% target).

Dataset	Target coverage	Empirical coverage	$\hat{q}$ (t/ha)	Avg width (t/ha)
Micro (t/ha)	0.9	0.9	0.8275	1.655
Macro (t/ha equiv.)	0.9	0.9009	4.441776170000001	8.88355156

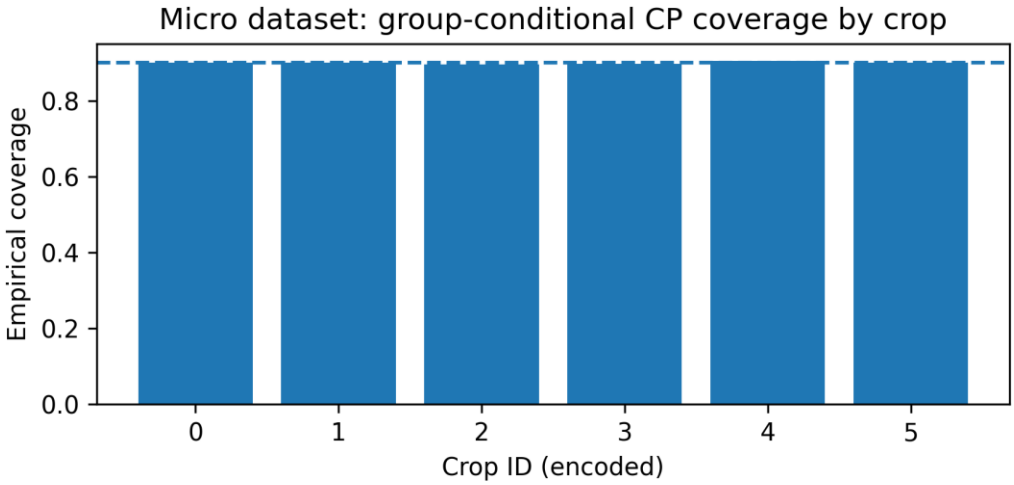

Figure. *6.* *Micro dataset: empirical 90% coverage by crop (group-conditional conformal).*

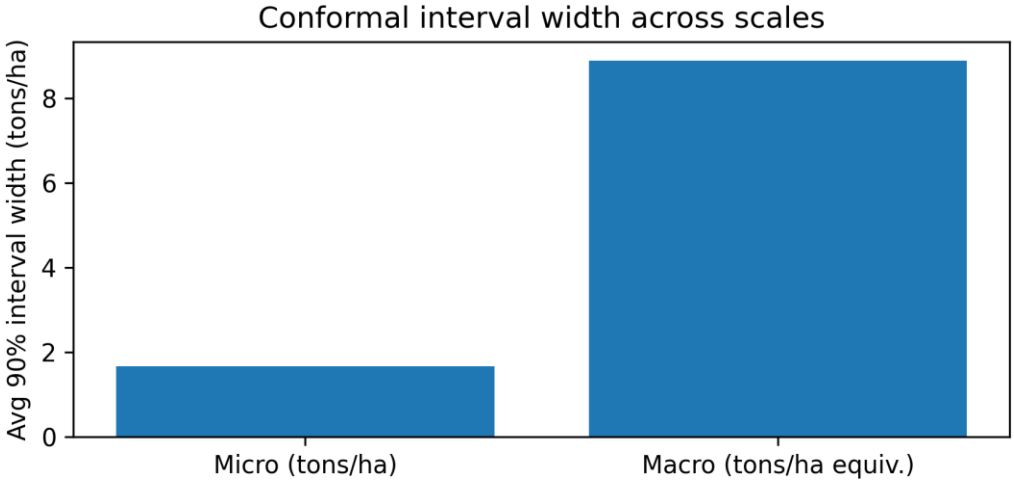

Figure. *7.* *Average conformal interval width across micro and macro scales (t/ha units).*

7. Discussion

The experiments provide several practical insights for cross-scale yield modeling. First, the synthetic micro dataset is highly learnable: linear regression, gradient boosting, and neural models achieve nearly identical performance. This suggests that the mapping from covariates to yield is close to additive after standardization and that complex

architectures deliver limited gains. In such settings, the main advantage of AgroX-MoE-CP is not a lower RMSE but the ability to produce calibrated prediction intervals through conformal calibration.

Second, the AgroCluster regime feature has limited marginal utility on the micro dataset. One plausible explanation is that the original categorical descriptors (Region, Soil_Type, Crop, Weather_Condition) already partition the data into relatively homogeneous groups, leaving little additional heterogeneity to be captured by KMeans over numeric variables. Another possibility is that the unsupervised clustering objective is misaligned with yield predictability; supervised regime discovery or representation learning could yield more predictive regime descriptors.

Third, the macro panel benefits from the same design pattern categorical embeddings plus MoE specialization when combined with log-stabilized regression. The macro dataset exhibits larger irreducible noise and structural heterogeneity (country/crop/technology trends), which translates to wider conformal intervals. Nevertheless, conformal prediction achieves the desired marginal coverage without assuming a parametric residual model, a property that is attractive under nonstationarity.

Fourth, group-conditional conformal calibration by crop type yields similar widths and overall coverage to global calibration on the micro dataset. This indicates limited residual heterogeneity across crops in the synthetic setting. In real-world datasets, subgroup-specific calibration may be more important, especially when some crops are underrepresented or have more variable responses to climate drivers.

A. Limitations

Several limitations should be noted. (i) The micro dataset is synthetic; while it enables large-scale benchmarking, it may underrepresent real-world complexities such as measurement error, missingness, management interactions, and spatial autocorrelation. (ii) The macro preprocessing standardizes numeric features on the full dataset before splitting. Although unsupervised, this step can reveal test distributional information and may slightly overestimate generalization performance; future implementations should fit scalers and clusters on the training data only. (iii) Our conformal intervals guarantee marginal coverage but not conditional coverage; under covariate shift or temporal drift, coverage may degrade. (iv) Macro-scale baselines are not exhaustively evaluated in this draft; future work should compare against strong tree ensembles and statistical panel models under consistent preprocessing.

B. Practical recommendations and future directions

Based on these findings, we recommend the following when deploying cross-scale yield models: (1) start with strong tabular baselines and treat complex architectures as augmentations rather than defaults; (2) prioritize uncertainty calibration (e.g., conformal prediction) when outputs feed into risk-sensitive decisions; (3) evaluate subgroup and time-slice coverage to identify failure modes.

Future research can extend AgroX-MoE-CP in several directions: (i) conditional and covariate-shift-aware conformal prediction to improve reliability under nonstationarity; (ii) asymmetric conformalized quantile regression for skewed residuals [6]; (iii) hierarchical or multi-level MoE where gating explicitly captures scale (plot vs country) and time; (iv) integration of remote-sensing imagery and temporal sequences via multi-modal transformers to leverage spatial context; and (v) interpretability analyses of gating and cluster assignments to connect learned regimes with agronomic knowledge.

8. Conclusion

This paper presented AgroX-MoE-CP, a cross-scale mixture-of-experts framework with conformal uncertainty calibration for crop yield prediction. The method combines unsupervised discovery of latent agro-climatic regimes (AgroClusters), a tabular MoE model with learned gating and expert specialization, and split conformal prediction to provide distribution-free prediction intervals. Across a synthetic micro plot-level dataset and a FAO-based macro panel, AgroX-MoE achieves strong point accuracy (micro: $R^2 \approx 0.91$; macro: $R^2 \approx 0.88$) and delivers 90% prediction intervals with empirical coverage ≈0.90. While point-accuracy gains over simple baselines are small on the synthetic micro data, the calibrated uncertainty is a key advantage for decision-making. Overall, the results support the feasibility of a unified, regime-aware, uncertainty-calibrated modeling recipe that is applicable across agricultural scales. Further work is needed to evaluate under real-world distribution shift and to integrate stronger macro baselines and multi-modal signals.

9. Acknowledgements

The authors thank the open-source community and dataset providers for making benchmarking resources publicly available. (Add funding and institutional acknowledgments here.)

This draft builds on conformal prediction foundations and predictive inference methods [7], [9], scalable clustering [10], open-source machine-learning frameworks [11], [12], optimization algorithms [13], [14], and strong tabular baselines such as gradient-boosted trees and random forests [15], [18]. For tabular deep learning, we draw on representative embedding-based and transformer-inspired models [19], [21], while the MoE design is motivated by both classic and modern sparse MoE architectures [22], [26]. Finally, we situate the work within crop-yield prediction and agricultural ML literature and data resources [27], [32].

IX. IMPLEMENTATION DETAILS AND ADDITIONAL RESULTS

This section provides implementation-oriented details and additional diagnostic tables to facilitate reproducibility and to support peer review.

A. Reproducibility checklist

- Fixed random seeds for Python, NumPy, and PyTorch.
- Four-way data split (train/val/calib/test) with crop stratification on the micro dataset.
- Standardization of numeric features and deterministic categorical encoding.
- Model selection using the validation set only; conformal calibration using a disjoint calibration set.

B. Model capacity and parameterization

Table VI reports the number of trainable parameters for two representative configurations used in this paper. Both models are lightweight compared to modern foundation models; this is intentional to preserve deployability in resource-constrained agricultural settings. Parameter counts exclude non-trainable buffers and depend on categorical cardinalities (especially for macro datasets).

Table VI. Model capacity (micro configurations).

Model variant	Experts	Hidden dim	Params (trainable)
AgroMoE-CP (baseline)	4	64	34666
AgroX-MoE (synthetic)	4	128	163564

C. Additional conformal diagnostics (micro)

Table VII reports crop-wise empirical coverage under group-conditional conformal prediction (α=0.1). Coverage is close to the nominal 0.90 target for all crop groups, indicating consistent uncertainty behavior across crops in the synthetic micro dataset.

Table VII. Micro dataset: empirical 90% coverage by crop (encoded IDs).

Crop ID (encoded)	Empirical coverage
0.0	0.901
1.0	0.901
2.0	0.896
3.0	0.898
4.0	0.905
5.0	0.9

D. Macro-unit conversion note

The macro dataset reports yield in hectograms per hectare (hg/ha). For cross-scale interpretability, we report a simple unit conversion to tons per hectare (t/ha): 1 hg/ha = 0.0001 t/ha. Table VIII summarizes the conversion used for macro metrics and interval widths.

Table VIII. Unit conversion used for macro-scale reporting.

Quantity	Equivalent
1 hectare (ha)	10,000 m²

1 hectogram (hg)	0.1 kg
hg/ha to t/ha	t/ha = (hg/ha) × 0.0001
Example (macro RMSE)	30103 hg/ha ≈ 3.010 t/ha

E. Algorithmic summary and complexity

Algorithm 1 summarizes the split conformal calibration procedure used in this paper. The approach is model-agnostic and adds a single post-training calibration pass of $O(|\ D_{\text{calib}}\ |)$to compute residual scores and the calibration quantile. At inference time, conformal prediction adds only constant overhead (two scalar operations per prediction). For the MoE model, a forward pass evaluates the gating network and all experts; with Eexperts and hidden dimension H, the dominant computation scales as $O(E \cdot H^2)$per sample for the expert MLPs, which is feasible for the lightweight settings used here ($E = 4$, $H \leq 128$).

Algorithm 1 Split conformal prediction (absolute residual score)
Input: training set D_train, calibration set D_calib, miscoverage α
Output: prediction interval PI(x) for any x
1: Train point model f on D_train
2: For each (x_i, y_i) in D_calib:
3: r_i ← | y_i − f(x_i) |
4: q̂ ← Quantile_{1−α}({r_i})
5: For any new x:
6: ŷ ← f(x)
7: PI(x) ← [ŷ − q̂ , ŷ + q̂]

In applications where residuals are skewed or heteroscedastic, Algorithm 1 can be adapted by using asymmetric scores or conformalized quantile regression [6], or by computing group-specific $\hat{q}$values (Mondrian conformal) to balance coverage across subpopulations.

X. EXTENDED CONTEXT: UNCERTAINTY AND DISTRIBUTION SHIFT IN AGRICULTURE

Yield prediction is often deployed under conditions that violate i.i.d. assumptions, including temporal drift (technology trends and climate change), covariate shift (new regions or extreme seasons), and label noise (reporting inconsistencies at the macro scale). While the present experiments use random splits, real deployments should include time-based evaluation and stress tests under extreme weather. Recent transformer-based and multi-modal approaches aim to improve robustness by leveraging spatial–temporal context and remote-sensing imagery; for example, MMST-ViT explicitly considers climate-change signals in yield prediction [30].

Uncertainty quantification (UQ) is essential in these settings: models should not only be accurate on average but should also know when they are likely to err. UQ methods vary in their assumptions and guarantees. Deep ensembles and Bayesian neural networks provide useful heuristics but do not guarantee coverage. Quantile regression offers conditional quantiles but may be miscalibrated without post-hoc correction. Conformal prediction provides a simple and principled alternative with finite-sample marginal coverage guarantees under exchangeability [5], and is increasingly explored for agricultural decision support [1].

A practical strategy is to pair a strong point predictor (tree ensemble or deep model) with a lightweight conformal calibration layer. This separation of concerns—accuracy from the predictor and reliability from calibration—supports modularity and can reduce engineering complexity. However, conformal intervals can still fail under severe distribution shift. In such cases, shift-aware conformal techniques, re-calibration with recent data, and monitoring of interval coverage over time are recommended.

Table IX provides a concise comparison of common UQ strategies in tabular yield modeling, highlighting the role of conformal prediction as a distribution-free reliability wrapper.

Table IX. Comparison of uncertainty quantification strategies for crop-yield regression.

Method	Uncertainty type	Coverage guarantee	Notes
Ensembles / bootstrapping	Epistemic + aleatoric (heuristic)	No	Requires multiple model fits; calibration may drift under shift.

Bayesian neural networks	Epistemic (model) + aleatoric (if modeled)	No	Requires approximate inference; sensitive to priors and misspecification.
Quantile regression	Conditional quantiles	No (without calibration)	Produces asymmetric intervals; calibration needed for finite-sample validity.
Split conformal prediction	Distribution-free predictive intervals	Yes (marginal)	Model-agnostic; interval width adapts to residual scale; can be group-conditional.

10. References

[1] M. Farag, A. Emam, J. Leonhardt, and R. Roscher, “Enhancing decision support in crop production: Analyzing conformal prediction for uncertainty quantification,” Computers and Electronics in Agriculture, vol. 237, pt. B, Oct. 2025, Art. no. 110559, doi:10.1016/j.compag.2025.110559.

[2] Md. A. Jabed and M. A. A. Murad, “Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability,” Heliyon, vol. 10, no. 17, Sep. 2024, Art. no. e40836, doi:10.1016/j.heliyon.2024.e40836.

[3] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, “A Survey on Mixture of Experts,” ACM Computing Surveys (to appear), 2024, arXiv:2407.06204, doi:10.48550/arXiv.2407.06204.

[4] S. Somvanshi, S. Das, S. A. Javed, G. Antariksa, and A. Hossain, “A Survey on Deep Tabular Learning,” arXiv:2410.12034, 2024, doi:10.48550/arXiv.2410.12034.

[5] A. N. Angelopoulos and S. Bates, “A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification,” arXiv:2107.07511, 2021.

[6] Y. Romano, E. Patterson, and E. J. Candès, “Conformalized Quantile Regression,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2019.

[7] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, “Predictive inference with the jackknife+,” Annals of Statistics, vol. 49, no. 1, pp. 486–507, 2021.

[8] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. Berlin, Germany: Springer, 2005.

[9] G. Shafer and V. Vovk, “A tutorial on conformal prediction,” Journal of Machine Learning Research, vol. 9, pp. 371–421, 2008.

[10] D. Sculley, “Web-scale k-means clustering,” in Proc. 19th Int. Conf. World Wide Web (WWW), 2010, pp. 1177–1178.

[11] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[12] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2019.

[13] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[14] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. Int. Conf. Learning Representations (ICLR), 2019.

[15] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 785–794.

[16] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[17] G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 3146–3154.

[18] L. Prokhorenkova et al., “CatBoost: Unbiased boosting with categorical features,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2018.

[19] S. O. Arik and T. Pfister, “TabNet: Attentive interpretable tabular learning,” arXiv:1908.07442, 2019.

[20] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, “Revisiting deep learning models for tabular data,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2021.

[21] N. Hollmann et al., “TabPFN: A transformer that solves small tabular classification problems in a second,” arXiv:2207.01848, 2022.

[22] N. Shazeer et al., “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” arXiv:1701.06538, 2017.

[23] W. Fedus, B. Zoph, and N. Shazeer, “Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity,” arXiv:2101.03961, 2021.

[24] C. Riquelme et al., “Scaling vision with sparse mixture of experts,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2021.

[25] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive mixtures of local experts,” Neural Computation, vol. 3, no. 1, pp. 79–87, 1991.

[26] M. I. Jordan and R. A. Jacobs, “Hierarchical mixtures of experts and the EM algorithm,” Neural Computation, vol. 6, no. 2, pp. 181–214, 1994.

[27] 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, 2020, Art. no. 105709.

[28] 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.

[29] A. Kamilaris and F. X. Prenafeta-Boldú, “Deep learning in agriculture: A survey,” Computers and Electronics in Agriculture, vol. 147, pp. 70–90, 2018.

[30] F. Lin et al., “MMST-ViT: Climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer,” arXiv:2309.05057, 2023.

[31] Food and Agriculture Organization of the United Nations, “FAOSTAT statistical database,” accessed: Jan. 2026.

[32] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in Proc. 5th Berkeley Symp. Math. Statist. Prob., 1967, pp. 281–297.