

Domain-Shift-Aware Climate-Smart Crop Portfolio Advisory for Nepal with Self-Supervised Tabular Learning and Conformal Shortlists

Lumeng Han¹, Derek Zhou^{*2}

Email: lmhan@ucdavis.edu

¹International Agricultural Development, University of California, Davis, CA, USA

²Computer Science, University of California, Berkeley, Berkeley, CA, USA

*Corresponding Author

Abstract

Crop advisory models can perform well on balanced benchmarks while failing under unfamiliar environmental or household conditions. This study evaluates two separate research tracks: crop-label classification using 2,200 agronomic benchmark records and observed crop-portfolio prediction using 43,753 village-census households and 981 TAFSSA households in Nepal. Eight crop classifiers were compared within a training pool, with independent temperature-calibration, conformal-calibration, and test subsets. Household models were evaluated under random and geographic holdouts, including 30 paired community-level ablations. An external explanation-consistency experiment used 900 held-out, human-annotated RAGTruth responses. Random forest attained the highest cross-validated macro-F1 (0.992), while LightGBM had the lowest conventional-model log loss (0.059). Temperature scaling produced negligible change in held-out log loss. At 90% nominal coverage, adaptive prediction sets included the benchmark label in 438 of 440 cases, with a mean of 1.09 crops. Environmental-cluster holdouts reduced macro-F1 to 0.000–0.656, partly because some test crops were absent from training. Village-census logistic-regression macro average precision declined from 0.680 under random partitioning to 0.578 under district holdout for the same test households. TAFSSA cross-district values were 0.332 and 0.445. A fixed-split contextual gain did not persist across community holdouts: full context reduced macro average precision by 0.037 relative to farm-only predictors. On RAGTruth, numeric checks detected 132 of 579 erroneous responses; a source-linked lexical screen detected 432 but also rejected 102 of 321 valid responses. The findings support cautious interpretation of benchmark shortlists and household associations. The components remain a research prototype without end-to-end or field validation of agricultural decisions in Nepal.

Keywords: Climate-Smart Agriculture; Nepal; Self-Supervised Tabular Learning; Conformal Prediction; Evidence Consistency.

I. INTRODUCTION

Climate-smart agriculture links productivity, adaptation, and mitigation, but an advisory system is useful only when its recommendation remains credible under local conditions. Nepalese farms differ in climate exposure, groundwater, machinery, landholding, infrastructure, and market access. The Food and Agriculture Organization (FAO) describes climate-smart agriculture as context specific (FAO, 2017), food-security investment is another central objective (Lipper et al., 2014), and the Intergovernmental Panel on Climate Change (IPCC) emphasizes heterogeneous exposure and vulnerability (IPCC, 2022). These objectives motivate this study; its experiments evaluate predictive reliability and observed crop portfolios rather than changes in yield, income, adaptation, or emissions.

Machine-learning crop recommenders often map soil nutrients and climate variables to one crop label. Random splits, however, assume that test cases resemble training cases; point predictions

hide uncertainty; and an agronomic label does not establish household feasibility. Smart-farming research stresses integration of heterogeneous data (Wolfert et al., 2017), yet many benchmarks still compress a decision into one label and one accuracy score. The first problem is domain shift. A model can perform well when each crop and environment appears in training and testing, yet fail when an environmental cluster or crop class is absent. Covariate-shift theory and evaluations across changing data distributions show why in-distribution performance can overstate field reliability (Koh et al., 2021; Sugiyama et al., 2007; Zhang, 2023). For crop advisory, a change in environmental predictors must also be distinguished from the absence of training examples for a requested crop.

The second problem is decision uncertainty. Calibration assesses whether probability estimates correspond to empirical correctness (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). Conformal prediction converts scores into candidate sets with target marginal coverage under exchangeability (Angelopoulos & Bates, 2023; Vovk et al., 2005), and adaptive sets expand for difficult cases (Romano et al., 2020). Selective prediction adds a reject option for low-confidence inputs (Geifman & El-Yaniv, 2019; Nie & Zheng, 2023). None of these mechanisms alone establishes that a crop will be suitable or profitable at a particular farm.

The third problem is explanation grounding. Retrieval-augmented generation can connect language outputs with source material (Lewis et al., 2020; Zhou et al., 2023), but fluent text may still misrepresent that material (Ji et al., 2023). Structured explanations offer a narrower alternative: they can report recorded inputs, prediction scores, and explicit qualifications in a controlled format. Their verification should distinguish a correct number or citation from a scientifically justified agronomic claim. This study uses a two-track design. Track A compares tabular classifiers and a paired self-supervised neural model on a controlled crop-label benchmark, then evaluates calibration, conformal shortlists, abstention, environmental holdouts, and low-label learning. Track B predicts crops observed in Nepal household portfolios using a village census and linked TAFSSA surveys. The two tracks are evaluated independently because their records do not share farm or plot identifiers.

A crop shortlist avoids forcing every case into one recommendation. In Track A, the prediction set concerns the benchmark crop label. In Track B, ranking concerns crops already observed under household and local conditions. Observed cultivation can reflect access constraints, tradition, or market conditions and is not a direct label of agronomic suitability. Both outputs therefore require a different interpretation from a prescription about what a household should plant.

The study asks which models balance classification and probability quality, how environmental and geographic holdouts change performance, whether contextual features help across repeated

community splits, and how numeric and lexical evidence checks perform on naturally occurring language-model errors. Its contribution is an evaluation design that separates these questions and tests their limits. Tree ensembles, masked representation learning, temperature scaling, adaptive prediction sets, and rule-based checking are established components; the methodological focus is the evaluation of domain shift, uncertainty, and evidence checks across the two separately assessed tracks.

II. LITERATURE REVIEW

A. *Climate-smart and data-driven crop advisory*

Climate-smart agriculture is location-dependent. Regional planning requires resource and implementation criteria as well as biophysical information (FAO, 2017; Lipper et al., 2014). This supports separating compatibility with soil and climate variables from household production patterns: neither crop labels nor observed portfolios establish that a farm can adopt a crop successfully. Agricultural machine learning covers yield estimation, disease detection, irrigation, and recommendation. Reviews report useful predictive results but identify data quality and transferability as persistent constraints, especially for small and medium agricultural tables (Kamilaris & Prenafeta-Boldú, 2018; Liakos et al., 2018). These constraints are particularly relevant when a balanced benchmark is used to motivate advice for a different country or farming system.

B. *Tabular models and self-supervised representations*

Tree ensembles remain strong tabular baselines. Random forests reduce variance through bootstrap aggregation (Breiman, 2001), while Extra Trees introduce additional split randomization (Geurts et al., 2006). Gradient boosting is another established family, including efficient implementations such as LightGBM and CatBoost (Ke et al., 2017; Prokhorenkova et al., 2018). Comparative work shows that tree models remain competitive with deep networks on typical medium-sized tables (Grinsztajn et al., 2022).

C. *Uncertainty and geographic validation*

scaling can improve or worsen individual held-out calibration measures (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). Log loss, expected calibration error, and the multiclass Brier score therefore complement classification metrics rather than being inferred from them (Zhong et al., 2023). Conformal prediction provides finite-sample marginal coverage under exchangeability, not universal protection across regions (Angelopoulos & Bates, 2023; Chen et al., 2024; Vovk et al., 2005). Adaptive prediction sets vary with case difficulty (Romano et al., 2020). A model fitted using one calibration subset must be held fixed when scores are computed on an independent

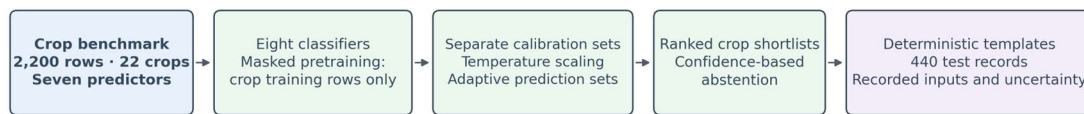
conformal subset. Explicit distribution-shift testing remains necessary because uncertainty estimates can deteriorate outside the training distribution (Koh et al., 2021; Ovadia et al., 2019).

Spatial validation changes the question from interpolation among familiar households to prediction in a new setting. Shared infrastructure and nearby farms can occur on both sides of a random household split, so community-based grouping is important (Roberts et al., 2017). Crop portfolios are also imbalanced. Precision–recall measures help assess rare outcomes (Saito & Rehmsmeier, 2015), but average precision depends on prevalence and should be interpreted with per-crop counts, ROC AUC, and simple prevalence baselines.

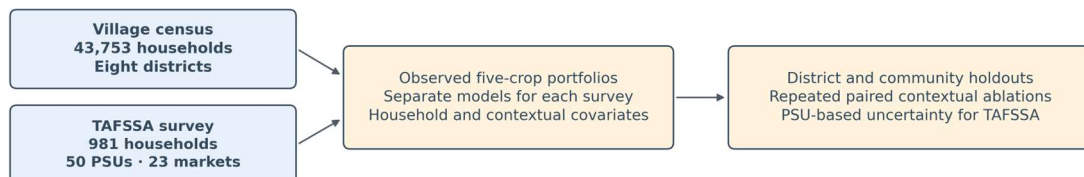
D. Structured explanations and verification

An explanation can preserve exact values while still making an unsupported causal or economic inference. Structured templates constrain the information presented but do not establish the correctness of an agricultural recommendation. Broader language-faithfulness research distinguishes fluent answers from grounded claims (Ji et al., 2023; Li, 2024). RAGTruth provides naturally generated responses with human hallucination annotations (Niu et al., 2024), allowing an external test of evidence checks beyond deliberately deleted identifiers or altered numbers.

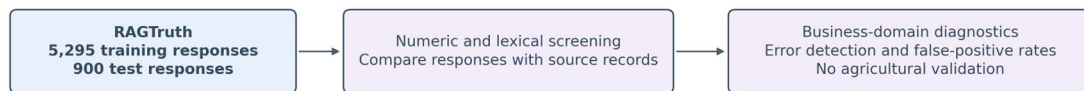
Track A | Crop classification and structured explanations



Track B | Independent analyses of observed household crop portfolios



External diagnostic | RAGTruth data-to-text benchmark



A household-to-crop interface remains conceptual; joint recommendations were not evaluated.

Figure 1. Research architecture. Crop classification, Nepal household-portfolio prediction, and external language-consistency testing are evaluated separately. The proposed advisory interface has no paired farm-level validation.

III. RESEARCH METHOD

A. Study design and data Architecture

Track A estimated benchmark crop labels from nitrogen, phosphorus, potassium, temperature, humidity, soil pH, and rainfall. Track B estimated crops observed in household portfolios from household and local-context variables. Figure 1 separates these independently evaluated components from the external language-consistency experiment. Table 1 lists the data sources and analytical units. The benchmark and household records do not share farm or plot identifiers, so their predictions were not combined into an experimentally validated recommendation.

Table 1. Data resources and analytical roles

Resource	Analytical sample	Role
Crop Recommendation (Ingle, 2020)	2,200 records; 22 crop labels	Crop classification, calibration, prediction sets and covariate-shift stress tests
Nepal Village Census (Krishna & Khed, 2024)	43,832 source rows; 43,753 households analyzed; 8 districts	Five observed cultivation outcomes and district holdouts
TAFSSA Male (IFPRI, 2024b)	1,000 source households; 981 analyzed; Banke and Surkhet	Five observed crop-portfolio outcomes
TAFSSA Community (IFPRI, 2024a)	50 linked primary sampling units (PSUs)	Infrastructure, groundwater, machinery and labor context
TAFSSA Market (IRRI & IFPRI, 2024)	23 markets; 2 district-level profiles	Market access and crop-price context
RAGTruth (Niu et al., 2024)	5,295 training and 900 test responses; 1,033 source records	External data-to-text evidence-consistency evaluation

B. Data preparation and outcome construction

The crop benchmark contains 100 records per label, seven complete numeric predictors, and no duplicate rows. Its provider describes augmentation of Indian rainfall, climate, and fertilizer data (Ingle, 2020). It is a controlled classification benchmark rather than a set of measured Nepal farm outcomes. Nutrient values were retained on their recorded scales because their physical units were not established. Temperature, humidity, pH, and rainfall followed the source definitions.

The village-census workbook contains 43,832 records. Twelve nonconsenting records and 23 records with unresolved crop outcomes or district identifiers were excluded. All 44 otherwise eligible records sharing 22 duplicated primary household identifiers were also excluded, leaving 43,753 unique households. Five outcomes represented wheat, rice, maize, mungbean, and lentil cultivation during the preceding 12 months. Missing recent-cultivation answers were treated as structural zeros only when the household explicitly reported no prior cultivation of that crop or no cultivation of any crop during the preceding five years. Explicit recent answers took precedence over broader screening responses. Eight district codes were linked to names through the workbook codebook.

Twenty-one census predictors described administrative location, household-head age and sex, caste group, tubewell access, generic machinery and herbicide use, coordinates, altitude, and land area with its unit and conversion-rate fields. Crop-specific fields used to construct outcomes were

excluded from predictors. Municipality and ward categories incorporated their parent administrative units. Zero or invalid coordinates were treated as missing. The land amounts remained on their reported scales, with unit information retained as covariates. Concurrent predictors and recent cultivation outcomes support retrospective associations rather than prospective adoption or causal effects.

TAFSSA household and community records were linked by their recorded household and PSU identifiers. The crop roster contained 3,152 records for 759 households. Rice, maize, lentil, chickpea, and blackgram indicators were positive when the crop appeared in any recorded season. A missing crop roster was assigned zero outcomes only when the land module explicitly reported no crop farming. Seventeen farming households without a roster and two households without a roster or a usable farming-status response were excluded. Explicit crop records took precedence for six households with a conflicting negative farming screen. The analytical sample contained 487 households in Banke and 494 in Surkhet, distributed across 50 PSUs.

The TAFSSA predictor blocks contained 12 household/farm variables, seven climate-information variables, 27 community variables, and nine market variables. Land amounts were converted to hectares using the stated units before operated land was calculated as owned plus leased-in minus leased-out area; ambiguous positive land-unit combinations were treated as missing (Central Bureau of Statistics, 2014). Machinery-rental comparisons used hourly quotations. Kilogram-denominated crop-price quotations were aggregated within market and then district; maize prices were unavailable. Thus the 23 markets supplied only two district-level contextual profiles. Household/PSU identifiers, production and sales outcomes, the farming screen, and crop-choice adaptation responses were excluded as predictors. Numeric imputation and standardization and categorical one-hot encoding were fitted within each training partition, with unseen categories ignored.

C. Crop models and self-supervised pretraining

The crop data were partitioned into a stratified training pool of 1,100 records, separate temperature- and conformal-calibration pools of 330 records each, and a held-out test set of 440 records. Five-fold stratified cross-validation within the training pool compared multinomial logistic regression, random forest, Extra Trees, histogram gradient boosting, LightGBM, CatBoost, a randomly initialized multilayer perceptron (MLP), and a masked-pretrained MLP. The conventional classifier with the lowest mean validation log loss was selected for the uncertainty experiments. The calibration and test records were excluded from model selection.

Random forest and Extra Trees each used 160 trees. Histogram gradient boosting used 180 iterations, learning rate 0.08, at most 31 leaf nodes, and L2 regularization of 0.05; LightGBM and

CatBoost used 220 estimators or iterations. Logistic regression used standardized inputs and L2 regularization with $C = 2$. The two MLPs shared a seven-input, 40-unit tanh encoder and a 22-class softmax head, identical initialization seeds, and the same supervised fitting and stopping rules. Masked pretraining reconstructed 35% randomly hidden predictor entries using only available crop-training covariates. The supervised validation portion was excluded from scaling and pretraining. Training used Adam optimization, with stopping based on internal validation loss. The implementation used NumPy for the neural models and scikit-learn for preprocessing and conventional estimators (Pedregosa et al., 2011).

This is self-supervised representation pretraining on a small tabular task. It does not use a broadly pretrained tabular foundation model. In low-label experiments, 10%, 25%, 50%, or 100% of the 1,100-record training pool provided labels, with three repeated runs and the same 440-record test set. Both MLPs used the same available unlabeled training covariates for scaling; only the masked-pretrained model additionally learned reconstruction weights. The external calibration and test records were never included in this pool.

D. Calibration conformal shortlists and abstention

Crop discrimination was measured by accuracy, macro-F1, and top-three accuracy. Probability quality was measured by log loss, 15-bin expected calibration error (ECE), and the multiclass Brier score. A scalar temperature minimized negative log likelihood on the temperature-calibration subset. The calibrated model was then held fixed when adaptive prediction-set scores were computed on the independent conformal subset. Each score combined cumulative probability above a candidate with a uniform random fraction of its probability (Guo et al., 2017; Romano et al., 2020).

Finite-sample higher quantiles defined sets at nominal coverage levels of 80%, 90%, and 95%. The same test randomization was used across levels to preserve nesting. An empty-set safeguard retained the highest-probability crop. This safeguard can increase empirical coverage and singleton frequency. Confidence thresholds for selective prediction were chosen on temperature-calibration records and applied unchanged to the test set. Accepted-case risk was the fraction of errors among accepted records. Exact binomial intervals describe observed test accuracy and prediction-set coverage; they do not establish farm-level validity or coverage under geographic shift (Angelopoulos & Bates, 2023).

E. Environmental holdouts and geographic validation

Standardized crop covariates were clustered into six environmental groups using k-means. Each cluster was held out while LightGBM was fitted to the remaining records. The partition used covariates from the entire benchmark and served as a descriptive stress test. Training sizes and

crop-label support differed across clusters. Accuracy and macro-F1 were therefore reported alongside counts of test crops absent from training and accuracy restricted to supported crop classes. These clusters were not geographic regions or an independent Nepal test set.

Household portfolios were modeled with five independent logistic classifiers and 160-tree random-forest and Extra Trees models. Random splits used 75% of eligible households for training and 25% for testing, with seed 42. Logistic regression used balanced class weights, L2 regularization with $C = 1$, and the liblinear solver. The forests used maximum depths of 14 and 16 and minimum leaf sizes of six and four, respectively. The same logistic specification was used for random and district-held-out comparisons. The census analysis withheld each of eight districts; TAFSSA transferred in both directions between Banke and Surkhet. Models predicted observed cultivation rather than the suitability of unobserved planting decisions.

Macro average precision (AP) was the primary portfolio metric because crop frequencies were unequal (Saito & Rehmsmeier, 2015). Macro ROC AUC, macro- and micro-F1, exact-subset accuracy, and top-two hit rate provided complementary summaries. Top-two hit used all evaluated households as its denominator, including households with none of the five target crops. Confidence intervals used 1,000 bootstrap samples of census district–municipality–ward clusters or TAFSSA PSUs while keeping model predictions fixed. Census districts contained only five to eleven administrative clusters, so district-specific intervals are descriptive. A paired census comparison evaluated random and spatial logistic predictions for the same 10,939 test households; training sizes and geographic support still differed between these protocols.

TAFSSA feature ablations used 30 paired, district-stratified PSU holdouts (Roberts et al., 2017). Each run withheld 13 PSUs, alternating six and seven PSUs per district, and trained on the remaining 37. All feature sets shared the same training and test households within each run. Farm-only, sequentially expanded context, district-only, and farm-plus-district models were compared. Standard deviations summarize partition variability. For mean AP and paired feature differences, district-stratified PSU bootstrap weights were shared across partitions and models. These intervals are conditional on the observed districts, partitions, and fitted models. AP was set to zero when a bootstrap replicate lacked positives for a crop, retaining the fixed five-crop average; AUC intervals used replicates with both classes represented.

F. Structured explanations and external evidence checks

The agricultural explanation interface used deterministic local templates. Each template reported the record identifier, seven recorded input fields, three ranked crop scores, conformal shortlist, and abstention status. It described scores as benchmark-label probabilities and avoided yield or profit claims. No language model, provider API, or generative decoding procedure was used for

these agricultural explanations. The templates were instantiated from the 440 held-out crop records using the 90% prediction set and the threshold selected for 90% calibration acceptance. Their constrained form does not establish semantic safety for unconstrained language-model outputs.

An external experiment evaluated evidence consistency using the Data-to-text portion of RAGTruth (Niu et al., 2024). The fixed February 2024 snapshot contained 6,198 responses based on 1,033 structured business records. After excluding three truncated training responses, the original source-disjoint split contained 5,295 training responses from 883 sources and 900 test responses from 150 sources. A response was positive when it contained at least one human-annotated hallucination span. The test set contained 579 positives and 321 negatives. No source identifier or exact source record occurred in both partitions.

The first screen flagged responses containing a normalized numeral not found in the associated source. It tested numerical support without interpreting meaning. The second screen used L2-regularized logistic regression on 20 source-linked features describing text length, lexical and bigram overlap, unsupported numbers, negation density, and punctuation. Five-fold source-grouped training validation selected regularization $C = 10$ by mean AP and a probability cutoff of 0.49 by pooled balanced accuracy. Source identifiers, generator identity, and annotation-derived variables were excluded from predictors. The fitted screen was evaluated once on the original test set. A flag-all baseline exposed the effect of the high positive prevalence.

Table 2. Five-fold crop-model performance within the 1,100-record training pool

Model	Accuracy (SD)	Macro-F1 (SD)	Log loss	ECE	Brier	Top-3
Random forest	0.992 (0.009)	0.992 (0.009)	0.088	0.067	0.028	1.000
Extra trees	0.990 (0.009)	0.990 (0.009)	0.137	0.107	0.042	1.000
CatBoost	0.988 (0.010)	0.988 (0.010)	0.076	0.053	0.023	1.000
LightGBM	0.986 (0.006)	0.986 (0.006)	0.059	0.015	0.025	1.000
Histogram gradient boosting	0.982 (0.006)	0.982 (0.006)	0.070	0.019	0.030	0.998
Scratch MLP	0.978 (0.011)	0.978 (0.011)	0.061	0.024	0.031	1.000
Masked-pretrained MLP	0.977 (0.009)	0.977 (0.009)	0.064	0.024	0.033	1.000
Logistic regression	0.971 (0.010)	0.971 (0.010)	0.211	0.133	0.080	0.998

Detection was assessed using precision, recall, specificity, balanced accuracy, MCC, AUROC, and AP. Confidence intervals resampled the 150 test sources, retaining their associated model responses, for 2,000 bootstrap replicates. Detection by human error category was measured at the response level: flagging a response did not necessarily locate its annotated span, and categories could overlap. RAGTruth supplies pre-existing outputs from several language models; this experiment evaluates those outputs and does not generate new agricultural advice. Its business-domain results are an external diagnostic of the screens rather than validation of Nepal agronomic recommendations.

G. Relationship between the two tracks

Track B did not reorder or remove Track A candidates in the experiments. A future interface could display household-portfolio scores next to a crop shortlist after harmonizing crop names. The census outcomes overlap with four of the benchmark labels; wheat has no corresponding Track A label. All five TAFSSA outcomes are represented in the benchmark. Crops outside a household model's vocabulary would have unknown household support, not a zero feasibility score. Removing crops from a conformal set would require an additional coverage evaluation. Establishing the value of any combined ranking would also require soil, climate, household context, and outcomes observed for the same farm or plot.

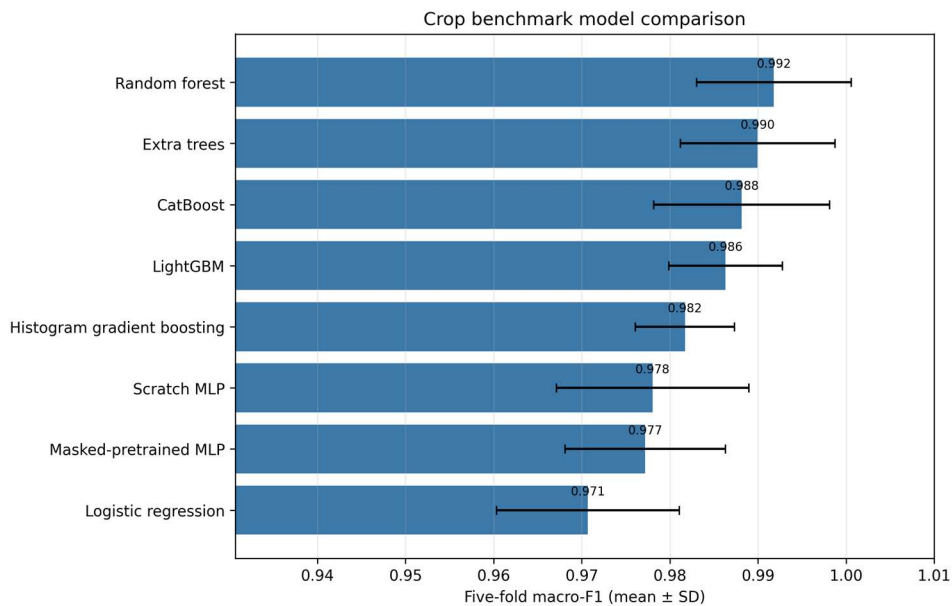


Figure 2. Cross-validated macro-F1 for the eight crop models. Error bars show one standard deviation across the five training-pool folds.

IV. RESULTS AND FINDINGS

A. Crop model comparison

Table 2 and Figure 2 compare models within the training pool. Random forest attained macro-F1 0.992, followed by Extra Trees at 0.990. LightGBM had the lowest mean log loss among the conventional models (0.059), which determined its selection for calibration and shortlist evaluation. The scratch and masked-pretrained MLPs attained macro-F1 0.978 and 0.977, respectively. This comparison did not show a general advantage for masked pretraining. Top-three accuracy was at least 0.998 across models, indicating that probability quality and distribution shift were more discriminating tests than inclusion among three benchmark labels.

Table 3. Held-out probability calibration

Method	Temperature	Accuracy	Log loss	ECE	Brier
Uncalibrated	1.0000	0.986	0.030609	0.010212	0.016138
Temperature scaled	1.0008	0.986	0.030615	0.010210	0.016136

B. Probability calibration and conformal shortlists

The fitted temperature was 1.0008. Both probability versions classified 434 of 440 test records correctly, giving accuracy 0.986 and a descriptive 95% exact interval of 0.971–0.995. Held-out log loss changed from 0.030609 to 0.030615, while ECE and Brier changes were negligible (Table 3; Figure 3). Temperature scaling therefore did not provide a meaningful calibration improvement in this partition.

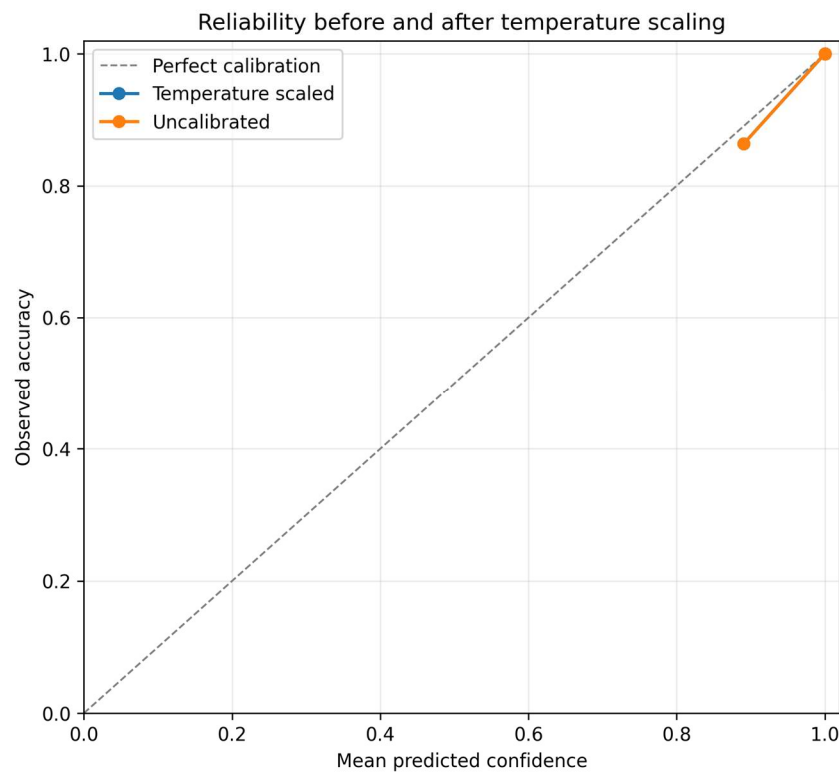


Figure 3. Reliability before and after temperature scaling on the crop test set. Near-overlapping curves are consistent with the fitted temperature close to one.

At 90% nominal coverage, the adaptive set contained the benchmark label in 438 of 440 records, or 99.55% (95% exact interval 98.37–99.94%), with a mean size of 1.09 crops (Table 4; Figure 4). The nonempty safeguard was applied to 44 cases at this level. Strong benchmark separation and the safeguard contributed to conservative coverage. The result concerns recovery of the recorded benchmark label, not the probability that every shortlisted crop would succeed on a Nepal farm.

Table 4. Randomized adaptive prediction sets on 440 test records

Nominal	Covered	Coverage % [95% CI]	Mean size	Singleton %	Empty-set safeguards
80%	437/440	99.32 [98.02, 99.86]	1.03	99.3	101
90%	438/440	99.55 [98.37, 99.94]	1.09	98.2	44
95%	440/440	100.00 [99.17, 100.00]	1.15	96.4	16

C. Selective Prediction

At full coverage, the test set contained six errors among 440 cases. A threshold chosen for 95% calibration coverage accepted 412 test cases and made no errors in that accepted subset (Table 5; Figure 5). Realized test coverage was 93.64%, illustrating that calibration coverage does not fix the acceptance rate of a new sample. Zero observed errors among 412 accepted cases corresponded to a 95% exact accuracy interval of 99.11–100%, rather than proof of zero future risk.

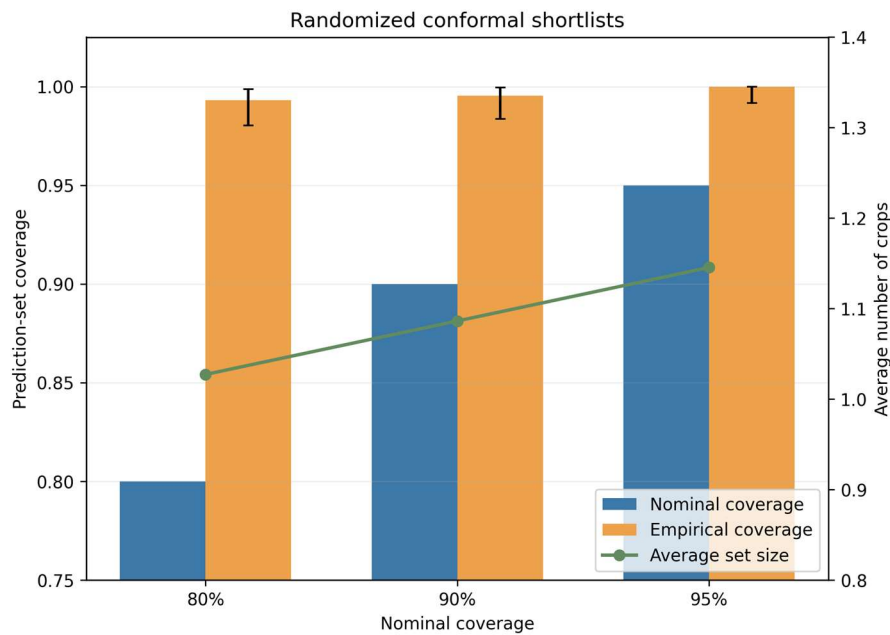


Figure 4. Nominal and empirical crop-label coverage and average prediction-set size.

Error bars show descriptive 95% exact intervals.

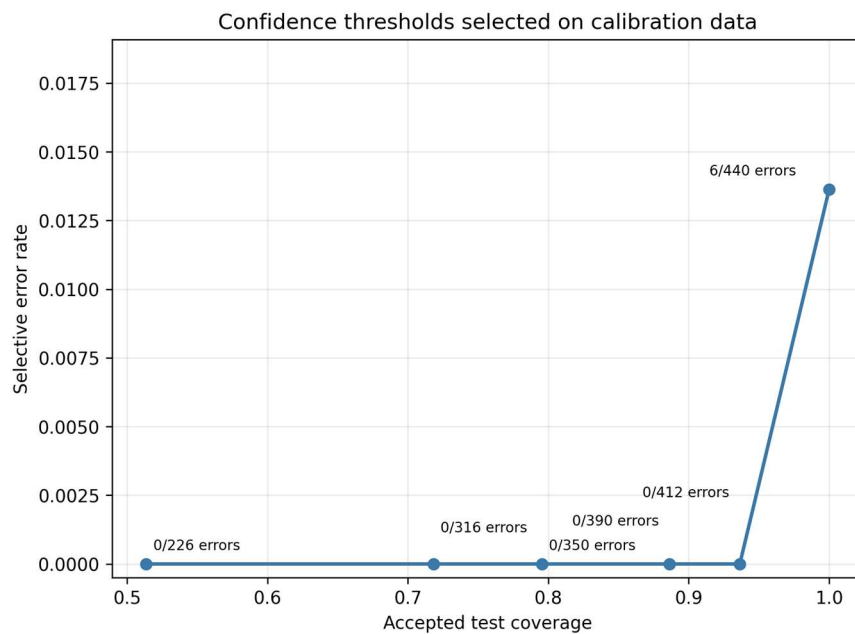


Figure 5. Selective error versus realized test coverage. Labels give error counts and accepted sample sizes; thresholds were selected on calibration records.

Table 5. Confidence-based selective prediction

Calibration target	Test coverage %	Threshold	Errors / accepted	Accuracy % [95% CI]
100%	100.00	0.000000	6/440	98.64 [97.06, 99.50]
95%	93.64	0.997153	0/412	100.00 [99.11, 100.00]
90%	88.64	0.999749	0/390	100.00 [99.06, 100.00]
80%	79.55	0.999945	0/350	100.00 [98.95, 100.00]
70%	71.82	0.999970	0/316	100.00 [98.84, 100.00]
50%	51.36	0.999983	0/226	100.00 [98.38, 100.00]

D. Environmental and crop-support shift

The six covariate clusters represented different nutrient and climate combinations (Table 6). Five holdouts removed at least one crop class from training. The apple–grape cluster, for example, contained 200 records and left neither label represented in the remaining data. These partitions combined changes in environmental covariates with loss of crop-label support.

Table 6. Mean predictor values in the six environmental clusters

Cluster	n	N	P	K	Temp °C	Humidity %	pH	Rain mm
0	457	27.6	54.8	22.3	29.1	72.2	7.10	65.7
1	566	49.9	32.8	34.6	24.6	85.4	6.52	168.9
2	556	99.5	41.4	37.5	25.6	78.6	6.44	75.7
3	200	22.0	133.4	200.0	23.2	87.1	5.98	91.1
4	198	20.2	45.6	24.5	29.9	49.8	5.42	102.8
5	223	29.5	67.7	46.9	19.5	22.3	6.50	95.8

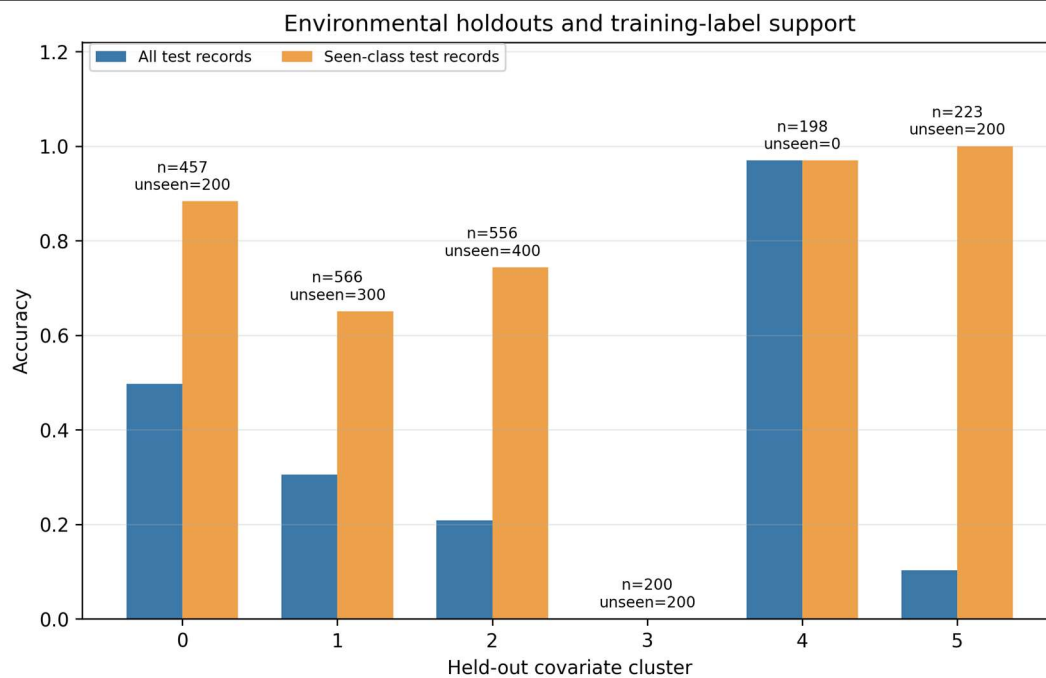


Figure 6. Accuracy under environmental-cluster holdout, including a separate view restricted to crop labels represented in training. Labels report test size and unseen-label count.

Holdout macro-F1 ranged from 0.000 to 0.656 (Table 7; Figure 6). The apple–grape holdout had zero accuracy because its test labels were absent from training. Restricting evaluation to seen crop

classes changed some results substantially: cluster 0 accuracy rose from 0.497 overall to 0.883 on supported labels. Conversely, cluster 1 supported-label accuracy remained 0.650. The results distinguish unsupported crop regimes from covariate shifts among known labels, while still reflecting different training sizes.

Table 7. LightGBM performance under environmental-cluster holdouts

Cluster	Test n	Unseen-label n	Accuracy	Macro-F1	Seen-class accuracy	Log loss
0	457	200	0.497	0.559	0.883	12.32
1	566	300	0.306	0.197	0.650	15.57
2	556	400	0.209	0.143	0.744	20.15
3	200	200	0.000	0.000	—	27.63
4	198	0	0.970	0.656	0.970	0.13
5	223	200	0.103	0.238	1.000	24.78
100%	Scratch MLP		1100	0.986 (0.004)		0.043

E. Low-label learning

Extra Trees retained macro-F1 0.972 with 10% of training-pool labels, compared with 0.871 for logistic regression, 0.851 for the scratch MLP, and 0.868 for the masked-pretrained MLP (Table 8; Figure 7). Masked pretraining produced small mean improvements over the matched scratch model in these repeated low-label evaluations, but the differences varied by fraction and were based on only three runs. At full training-pool size the two MLPs reached 0.986 and 0.988, respectively. These results support a limited initialization comparison rather than a claim of foundation-model or few-shot superiority.

Table 8. Low-label crop classification over three repeated runs

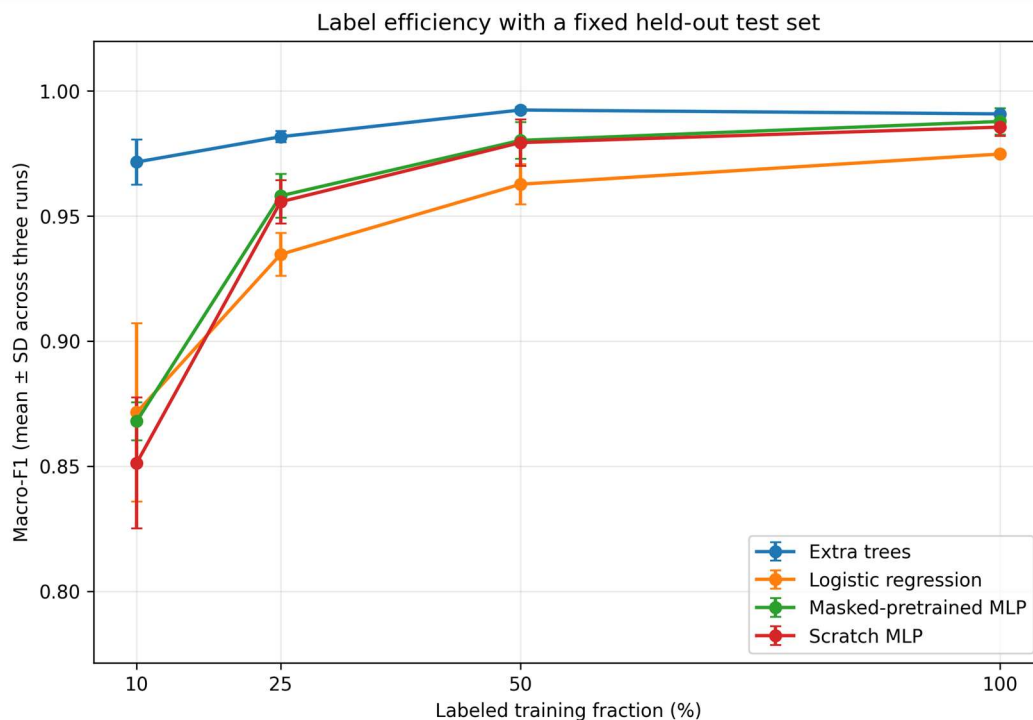
Labels %	Model	Training n	Macro-F1 (SD)	Log loss
10%	Extra trees	110	0.972 (0.009)	0.445
10%	Logistic regression	110	0.871 (0.036)	0.677
10%	Masked-pretrained MLP	110	0.868 (0.008)	0.457
10%	Scratch MLP	110	0.851 (0.026)	0.476
25%	Extra trees	275	0.982 (0.002)	0.276
25%	Logistic regression	275	0.935 (0.009)	0.417
25%	Masked-pretrained MLP	275	0.958 (0.009)	0.139
25%	Scratch MLP	275	0.956 (0.009)	0.141
50%	Extra trees	550	0.992 (0.001)	0.185
50%	Logistic regression	550	0.963 (0.008)	0.277
50%	Masked-pretrained MLP	550	0.980 (0.007)	0.068
50%	Scratch MLP	550	0.979 (0.009)	0.073
100%	Extra trees	1100	0.991 (0.002)	0.130
100%	Logistic regression	1100	0.975 (0.000)	0.197
100%	Masked-pretrained MLP	1100	0.988 (0.005)	0.044

F. Nepal village-census portfolios

Crop prevalence was unequal in both Nepal samples (Table 9). Rice and wheat were common in the village census, while mungbean was observed in 1,299 households. TAFSSA contained only 29 blackgram-growing households. These imbalances limit the precision of rare-crop estimates and make prevalence-sensitive measures such as AP essential to interpretation.

Table 9. Observed cultivation in the analytical household samples

Dataset	Crop	Households	Positive	Prevalence %
Village census	wheat	43,753	22,659	51.8
Village census	rice	43,753	26,707	61.0
Village census	maize	43,753	10,566	24.1
Village census	mungbean	43,753	1,299	3.0
Village census	lentil	43,753	14,052	32.1
TAFSSA	rice	981	675	68.8
TAFSSA	maize	981	451	46.0
TAFSSA	lentil	981	313	31.9
TAFSSA	chickpea	981	191	19.5
TAFSSA	blackgram	981	29	3.0

**Figure 7. Macro-F1 across labeled-data fractions. Error bars show one standard deviation across three runs on the same held-out test set.**

Random forest achieved the highest census random-split macro AP, 0.716 (95% conditional cluster-bootstrap interval 0.690–0.740), followed by Extra Trees at 0.693 and logistic regression at 0.680 (Table 10). The random-forest top-two list included an observed target crop for 6,891 of 10,939 households. Because 4,014 test households had none of the five target crops, the highest possible all-household hit rate was 63.31%. Top-two hit should therefore be read together with its denominator and AP rather than interpreted as an unrestricted recommendation success rate.

Table 10. Village-census random-split portfolio prediction

Model	Macro AP [95% CI]	Macro AUC	Macro-F1	Exact subset	Top-2 hit
Logistic one-vs-rest	0.680 [0.653, 0.704]	0.898	0.659	0.420	6,814/10,939
Random forest	0.716 [0.690, 0.740]	0.919	0.688	0.547	6,891/10,939
Extra trees	0.693 [0.667, 0.717]	0.892	0.635	0.403	6,889/10,939

Across the eight district holdouts, logistic-regression macro AP ranged from 0.472 in Rupandehi to 0.641 in Kailali (Table 11; Figure 8). Pooled district-held-out AP was 0.574. For the same 10,939 households used in the random test, spatial AP was 0.578 versus random AP 0.680, a difference of 0.101 (95% paired cluster-bootstrap interval 0.076–0.124). This comparison holds the evaluation households and model family constant, although the training sample sizes and geographical composition differ. It documents sensitivity to these eight sampled districts without establishing nationwide performance.

Table 11. Village-census district holdouts using logistic regression

District	Test n	Macro AP [95% CI]	Macro AUC	Macro-F1	Top-2 hit %
Banke	5,325	0.584 [0.548, 0.617]	0.866	0.563	56.4
Bardi	4,539	0.590 [0.523, 0.646]	0.855	0.567	59.3
Dang	6,404	0.617 [0.577, 0.649]	0.820	0.582	55.2
Kailali	8,482	0.641 [0.617, 0.659]	0.869	0.577	54.4
Kanchanpur	3,708	0.630 [0.551, 0.669]	0.874	0.513	54.0
Kapilwastu	6,324	0.554 [0.527, 0.586]	0.792	0.557	61.5
Parasi	3,046	0.563 [0.510, 0.577]	0.819	0.564	67.3
Rupandehi	5,925	0.472 [0.462, 0.489]	0.780	0.480	69.9

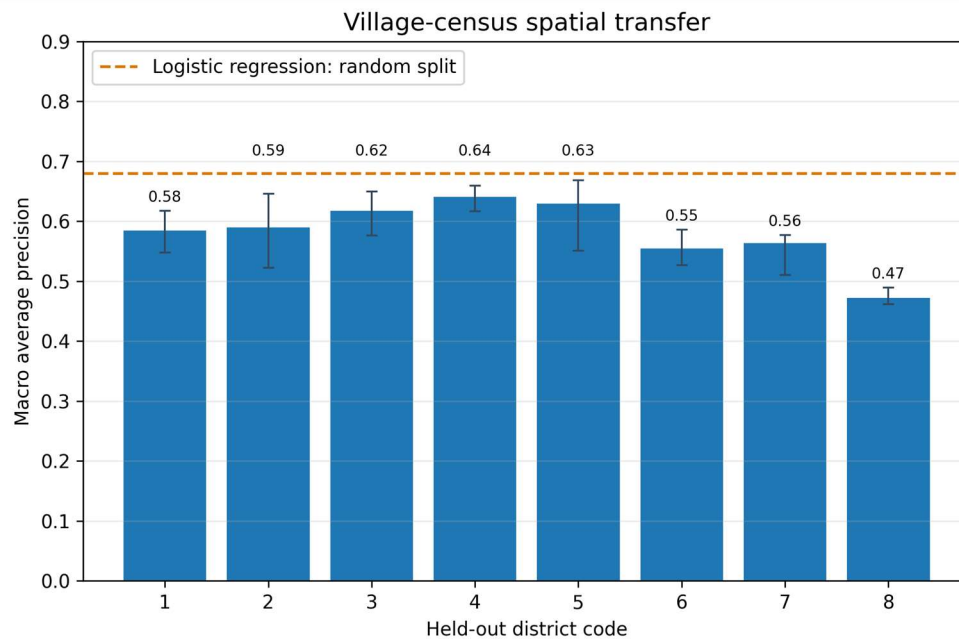


Figure 8. Village-census macro AP under district holdout. The reference is the same logistic model under a household-random split; error bars are conditional cluster-bootstrap intervals.

G. TAFSSA geographic transfer and contextual ablation

On the TAFSSA random test, Extra Trees reached macro AP 0.698, random forest 0.691, and logistic regression 0.646. With the same logistic model, training in Surkhet and testing in Banke yielded AP 0.332 (95% PSU-bootstrap interval 0.263–0.427); the reverse transfer yielded 0.445 (0.410–0.497). Macro AUC was 0.478 in Banke and 0.618 in Surkhet (Table 12; Figure 9). The

asymmetry and loss relative to random partitioning support a narrow conclusion about transfer between these two districts.

Table 12. TAFSSA household-random and district-held-out prediction

Evaluation	Model	Test n	Macro AP	Macro AUC	Macro-F1	Top-2 hit
Random	Logistic OVR	246	0.646	0.843	0.613	0.764
Random	Random forest	246	0.691	0.846	0.650	0.780
Random	Extra trees	246	0.698	0.827	0.623	0.780
Hold-out Bank	Logistic OVR	487	0.332	0.478	0.312	0.682
Hold out Surkhet	Logistic OVR	494	0.445	0.618	0.428	0.749

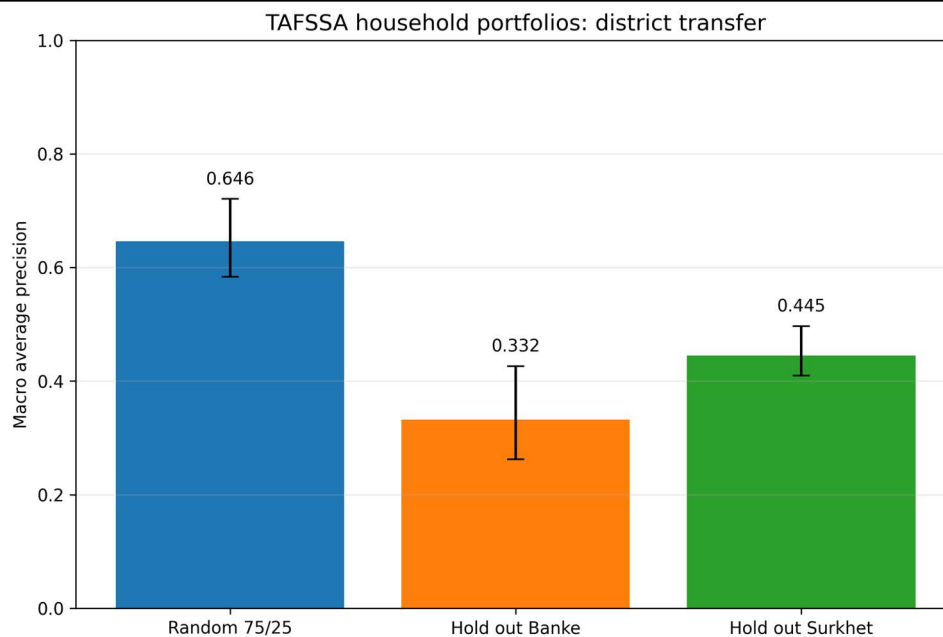


Figure 9. TAFSSA macro AP for the same logistic model under household-random and two directional district holdouts. Error bars show conditional PSU-bootstrap intervals.

Under the fixed random partition, full context raised macro AP from 0.582 to 0.646. This pattern did not persist when whole communities were withheld. Across 30 paired PSU holdouts, farm-only AP averaged 0.551 (SD 0.024), compared with 0.513 (SD 0.032) for full context (Table 13; Figure 10). The mean paired difference was -0.037 (95% conditional interval -0.055 to -0.012); full context improved performance in only three of 30 partitions. The community increment was -0.034 (-0.053 to -0.008), and the subsequent market increment was -0.0001 (-0.0027 to 0.0022).

Table 13. TAFSSA contextual ablation under random and community-held-out evaluation

Features	Count	Fixed-split AP	30-split AP (SD)	95% interval
Farm	12	0.582	0.551 (0.024)	[0.516, 0.581]
Farm + climate/info	19	0.582	0.548 (0.022)	[0.513, 0.579]
Farm + climate/info + community	46	0.646	0.513 (0.031)	[0.476, 0.554]
Full context	55	0.646	0.513 (0.032)	[0.476, 0.554]
District only	1	0.380	0.375 (0.040)	[0.326, 0.427]
Farm + district	13	0.585	0.554 (0.024)	[0.520, 0.584]

The district-only model averaged AP 0.375, while farm plus district averaged 0.554. Market features varied only between the two districts, so they cannot support a separate claim about market mechanisms. The fixed-split gain is consistent with local-context interpolation and did not demonstrate stable benefit for unseen PSUs under these model settings.

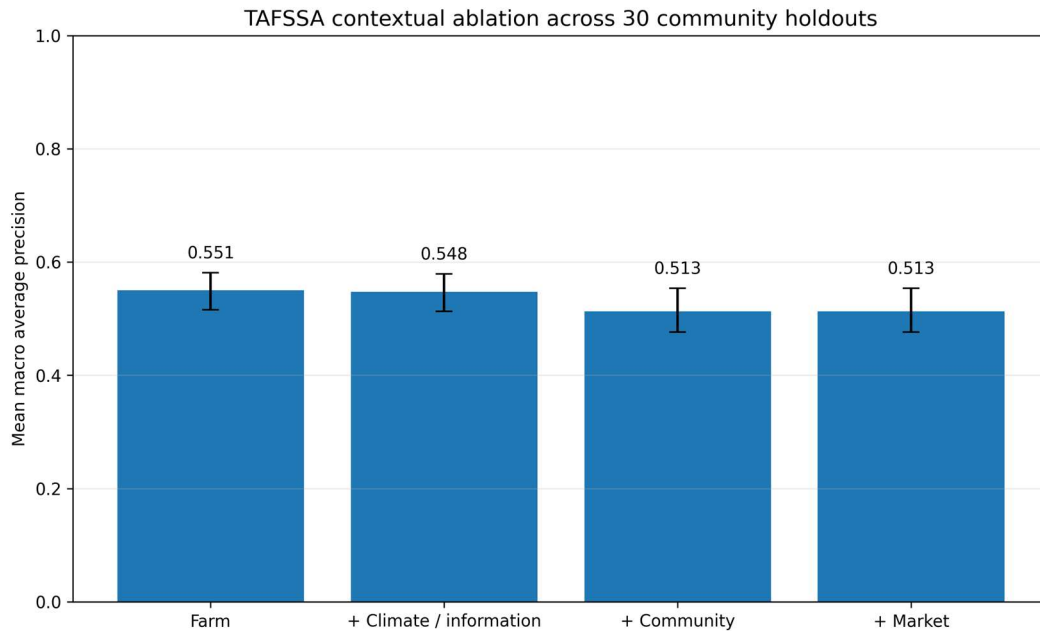


Figure 10. Contextual feature ablation across 30 paired PSU holdouts. Error bars show conditional district-stratified PSU-bootstrap intervals.

H. External explanation-consistency evaluation

The held-out RAGTruth responses contained 579 hallucination-positive and 321 negative cases. The numeric screen detected 132 positives but missed 447, giving recall 0.228 and specificity 0.810. The source-linked lexical screen detected 432 positives and missed 147, while rejecting 102 valid responses (Table 14; Figure 11). Its recall was 0.746 (95% source-bootstrap interval 0.710–0.782), specificity 0.682 (0.623–0.738), balanced accuracy 0.714 (0.681–0.747), AUROC 0.769 (0.735–0.802), and AP 0.832 (0.797–0.868).

Table 14. Evidence-consistency screens on 900 held-out RAGTruth responses

Screen	TP / FN	TN / FP	Precision	Recall	Specificity	Balanced accuracy
Numeric support	132/447	260/61	0.684	0.228	0.810	0.519
Lexical logistic	432/147	219/102	0.809	0.746	0.682	0.714
Flag all	579/0	0/321	0.643	1.000	0.000	0.500

The lexical screen's F1 of 0.776 was slightly below the flag-all baseline's 0.783 because hallucinations were common in this test set. The baseline had zero specificity, however, whereas the lexical screen retained 219 of 321 valid responses. This trade-off is better represented by balanced accuracy, false rejections, and the confusion counts than by F1 alone. Neither screen established robust semantic grounding.

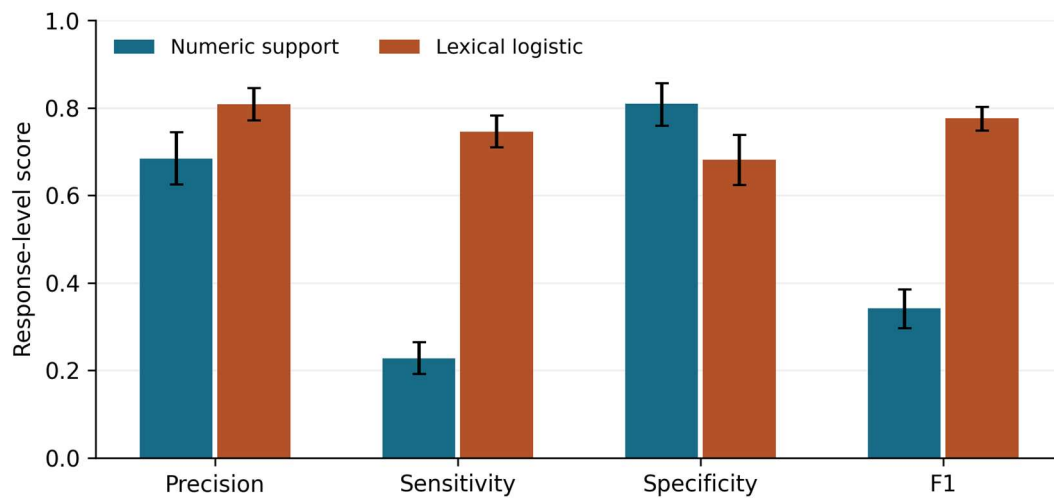


Figure 11. Precision, recall, specificity, and F1 of numeric and source-linked lexical screens on RAGTruth. Error bars resample test sources and their associated responses.

Figure 12 examines responses containing evident or subtle conflicts and unsupported information. Numeric consistency alone missed errors that did not introduce new numbers. A flagged response could also contain several error types, so these rates measure case detection rather than correct identification of each erroneous span. The subtle-conflict group contained only five responses, which is insufficient for a general robustness claim. The business-domain evaluation reveals limitations of these checks across naturally generated and human-annotated errors, but it does not test the agronomic validity of crop advice.

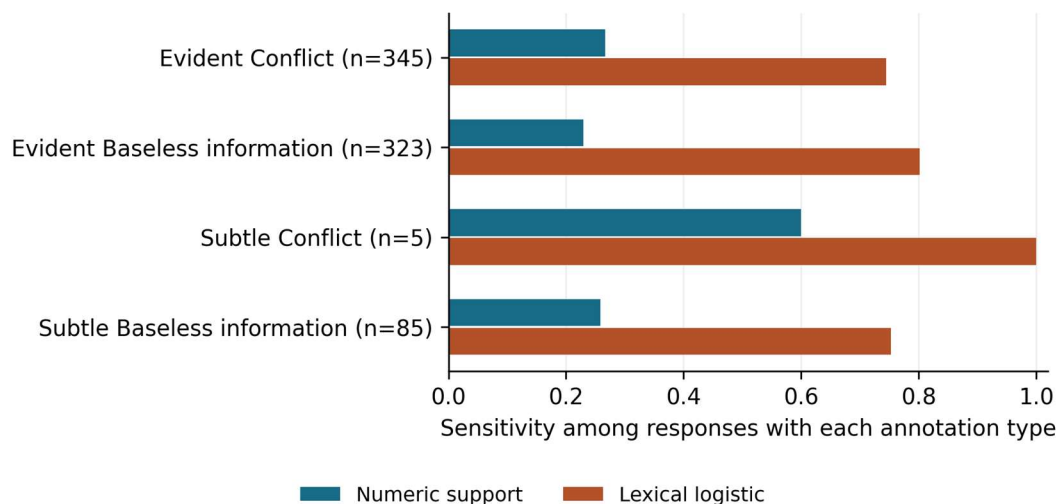


Figure 12. Response-level detection by RAGTruth human annotation category. Categories overlap; a response-level flag does not establish localization of its annotated error

I. Discussion

The strongest finding is the difference between interpolation and transfer. Crop classifiers performed well on the balanced benchmark, but environmental holdouts sometimes removed

entire crop labels from training. In Nepal, the same logistic model lost AP under district holdout, and TAFSSA context gains under household-random partitioning reversed when PSUs were withheld. These results support geographic and class-support checks as evaluation priorities, rather than a broad claim that a high random-split score transfers to new farms.

The community result also changes how contextual features should be interpreted. More variables increased discrimination among households drawn from familiar communities, but did not improve the tested model in unseen PSUs. Shared infrastructure and district-level market profiles can encode local patterns whose relationship to crops changes across locations. The two-district TAFSSA design cannot estimate nationwide market effects, and repeated resampling within these districts does not create new geographical evidence.

The three evidence levels should remain distinct. Track A assesses classification and uncertainty for an augmented crop-label benchmark. Track B assesses retrospective household crop portfolios. The RAGTruth experiment assesses detection of annotated errors in general-domain text. A good result in any one component does not establish the others, and no experiment evaluates a linked soil-to-household-to-advice decision on the same farm. Observed cultivation is influenced by constraints and preferences; it is not a counterfactual label of what should be planted.

The uncertainty results require similar restraint. Temperature scaling made almost no difference in this partition. Adaptive sets were conservative partly because empty sets were replaced by the top crop, and their coverage concerns the recorded label. Abstention removed all observed errors at several thresholds, but the accepted samples were finite and came from the same benchmark distribution. Neither set coverage nor zero observed selective errors establish safety under Nepal field conditions.

The explanation results show why preserving a number is insufficient. A response can use source-supported numbers while reversing a relationship, adding an unsupported cause, or implying a guarantee. The lexical screen improved recall over numeric checking but incurred substantial false rejection and missed 147 positive responses. Its source-disjoint evaluation is useful for identifying this limitation, although its business records and model-output styles differ from agricultural advice. The deterministic crop templates constrain wording; their factual scope remains limited to input records and model outputs.

The contribution is the combination of explicit domain-shift testing, independent uncertainty calibration, grouped contextual ablation, and bounded explanation evaluation. The underlying classifiers, masked reconstruction, temperature scaling, conformal methods, and lexical screening are established techniques. Additional seasons, representative regions, paired farm measurements,

and independent agronomic assessment would be necessary to evaluate adoption outcomes or deployment.

V. CONCLUSION AND RECOMMENDATION

This study evaluated a two-track crop advisory research prototype using a crop-label benchmark and Nepal household surveys, with a separate external test of language-output consistency. Tree ensembles performed strongly within the benchmark, but environmental holdouts exposed both covariate sensitivity and absent crop-label support. Calibration produced negligible change, while conformal shortlists and abstention summarized uncertainty within the evaluated distribution.

The Nepal experiments support geographically limited conclusions. Village-census prediction weakened under district holdout, and TAFSSA transfer was assessed only between Banke and Surkhet. Contextual gains from a household-random split did not persist across community holdouts. Simple evidence checks also missed many human-annotated hallucinations, even when a learned lexical screen improved sensitivity. These components have been evaluated as research methods, not as an end-to-end or agronomically validated decision system for Nepalese deployment.

Further development should prioritize local class support, repeated geographic and seasonal validation, clear presentation of uncertain or unsupported candidates, and agronomist review. Testing whether household context should reorder a crop shortlist requires paired farm-level observations and a separate evaluation of recommendation quality and coverage. The current evidence supports these research priorities without establishing yield, profit, or adoption benefits.

DATA AVAILABILITY STATEMENT

The crop benchmark is available on Kaggle (Ingle, 2020), and the village census is available from the CIMMYT Research Data repository (Krishna & Khed, 2024). The TAFSSA household, community, and market surveys are available through Harvard Dataverse (IFPRI, 2024a; IFPRI, 2024b; IRRI & IFPRI, 2024). RAGTruth is available from its authors' public repository (Niu et al., 2024); the analysis used commit 1d52a81c9e28e79e252a1945d858eb8dfd975c23. Data-source identifiers and publication years are provided in the reference list. The computational supplement contains analysis scripts, source checksums, partitions, predictions, metric tables, and figure-generation code.

AI USAGE AND GENERATIVE AI DECLARATION

OpenAI ChatGPT assisted software development, computational analysis, visualization, and manuscript drafting and editing. The agricultural explanations were generated with deterministic templates rather than a language model or generation API. The external RAGTruth experiment

analyzed pre-existing model responses and human annotations. Generative AI is not an author; responsibility for the submitted work rests with the authors.

REFERENCES

- Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494–591. <https://doi.org/10.1561/22000000101>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Central Bureau of Statistics. (2014). Statistical pocket book Nepal 2014. Government of Nepal, National Planning Commission Secretariat. https://giwmscdnone.gov.np/media/app/public/36/posts/1695537893_62.pdf
- Chen, S., He, S., & Sun, E. (2024). Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters. *Journal of Advanced Computing Systems*, 4(5), 119–134. <https://doi.org/10.69987/JACS.2024.40509>
- Food and Agriculture Organization of the United Nations. (2017). Climate-smart agriculture sourcebook (2nd ed.). <https://www.fao.org/climate-smart-agriculture-sourcebook/>
- Geifman, Y., & El-Yaniv, R. (2019). SelectiveNet: A deep neural network with an integrated reject option. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 2151–2159). PMLR. <https://proceedings.mlr.press/v97/geifman19a.html>
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63, 3–42. <https://doi.org/10.1007/s10994-006-6226-1>
- Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35, 507–520. <https://arxiv.org/abs/2207.08815>
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. <https://proceedings.mlr.press/v70/guo17a.html>
- Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2023). TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=cp5PvcI6w8_
- Ingle, A. (2020). Crop recommendation dataset (Version 1) [Data set]. Kaggle. <https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset>
- Intergovernmental Panel on Climate Change. (2022). Climate change 2022: Impacts, adaptation and vulnerability. Cambridge University Press. <https://doi.org/10.1017/9781009325844>
- International Food Policy Research Institute. (2024a). TAFSSA district agrifood systems assessment in Nepal 2023: Community (Version 1) [Data set]. Harvard Dataverse. <https://doi.org/10.7910/DVN/PZMNBU>

- International Food Policy Research Institute. (2024b). TAFSSA district agrifood systems assessment in Nepal 2023: Male [Data set]. Harvard Dataverse. <https://doi.org/10.7910/DVN/Z0RF6L>
- International Rice Research Institute, & International Food Policy Research Institute. (2024). TAFSSA district agrifood systems assessment in Nepal 2023: Market survey (Version 1) [Data set]. Harvard Dataverse. <https://doi.org/10.7910/DVN/POVL6D>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248. <https://doi.org/10.1145/3571730>
- Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70–90. <https://doi.org/10.1016/j.compag.2018.02.016>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154. <https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html>
- Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balasubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., . . . Liang, P. (2021). WILDS: A benchmark of in-the-wild distribution shifts. *Proceedings of Machine Learning Research*, 139, 5637–5664. <https://proceedings.mlr.press/v139/koh21a.html>
- Krishna, V., & Khed, V. (2024). Farmer census: Adoption of sustainable intensification technologies in Nepal Terai region (Version 1) [Data set]. CIMMYT Research Data & Software Repository Network. <https://hdl.handle.net/11529/10548922>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- Li, Y. (2024). Findable then explainable: Retrieval–summary integration for code intelligence on a lightweight CodeSearchNet subset. *Journal of Advanced Computing Systems*, 4(7), 65–82. <https://doi.org/10.69987/JACS.2024.40706>
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8), Article 2674. <https://doi.org/10.3390/s18082674>
- Lipper, L., Thornton, P., Campbell, B. M., Baedeker, T., Braimoh, A., Bwalya, M., Caron, P., Cattaneo, A., Garrity, D., Henry, K., Hottle, R., Jackson, L., Jarvis, A., Kossam, F., Mann, W., McCarthy, N., Meybeck, A., Neufeldt, H., Remington, T., . . . Torquebiau, E. F. (2014). Climate-smart agriculture for food security. *Nature Climate Change*, 4, 1068–1072. <https://doi.org/10.1038/nclimate2437>

- Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (pp. 625–632). Association for Computing Machinery. <https://doi.org/10.1145/1102351.1102430>
- Nie, J., & Zheng, D. (2023). Ambiguity-aware HDFS log anomaly detection with retrieval-augmented failure narratives and selective refusal. *Journal of Advanced Computing Systems*, 3(1), 66–80. <https://doi.org/10.69987/JACS.2023.30105>
- Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., & Zhang, T. (2024). RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 10862–10878). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.585>
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32, 13991–14002. <https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://jmlr.org/papers/v12/pedregosa11a.html>
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31, 6638–6648. <https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html>
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Aroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. <https://doi.org/10.1111/ecog.02881>
- Romano, Y., Sesia, M., & Candès, E. J. (2020). Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33, 3581–3591. <https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html>
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), Article e0118432. <https://doi.org/10.1371/journal.pone.0118432>

- Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8, 985–1005. <https://jmlr.org/papers/v8/sugiyama07a.html>
- Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer. <https://doi.org/10.1007/b106715>
- Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M.-J. (2017). Big data in smart farming: A review. *Agricultural Systems*, 153, 69–80. <https://doi.org/10.1016/j.agsy.2017.01.023>
- Zhang, H. (2023). DriftGuard: Multi-signal drift early warning and safe re-training/rollback for CTR/CVR models. *Journal of Advanced Computing Systems*, 3(7), 24–40. <https://doi.org/10.69987/JACS.2023.30703>
- Zhong, Z. S., Ma, R., & Zhao, H. (2023). Human-uncertainty distillation for calibrated vision models on CIFAR-10H. *Journal of Advanced Computing Systems*, 3(2), 77–89. <https://doi.org/10.69987/JACS.2023.30206>
- Zhou, S., Li, Z., & Wang, E. (2023). Evidence-grounded RAG for tokenized trade receivable disclosure QA under U.S. capital market standards. *Journal of Advanced Computing Systems*, 3(7), 41–57. <https://doi.org/10.69987/JACS.2023.30704>